THE UNIVERSITY
of EDINBURGH

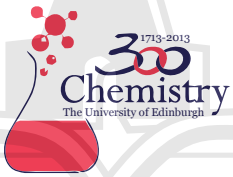
A dissertation submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy at the University of Edinburgh

Accelerating Molecular Simulations Implication for Rational Drug Design

by

Gaetano Calabrò

School of Chemistry
University of Edinburgh



October 2015

Supervisor: Dr Julien Michel

Declaration of Authorship

I hereby certify that this thesis has been composed by me and is based on my own work, unless stated otherwise. No other person's work has been used without due acknowledgement in this thesis. All references and verbatim extracts have been quoted, and all sources of information, including graphs and data sets, have been specifically acknowledged.

Edinburgh, 19^{th} October 2015

Gaetano Calabrò

Acknowledgement

I am indebted to the many people who made this thesis possible. First and foremost, I would like to thank my gratitude to my supervisor Dr Julien Michel, whose, expertise, understanding, and patience, added considerably to my PhD experience. I am indebted with him for many of my computational skills, for engaging me with new ideas and demanding a high quality of work in all my endeavors.

A very special thanks goes out to Dr Christopher Woods for providing an invaluable support and patiently taught me the Sire programming.

I am very grateful to my colleagues Remì and George, many times I saw their smiles blow away clouds from the grey skies of Edinburgh and from my mind. I must also acknowledge all the other group members Harris, Juan, Stefano, Kevin and Pattama.

I also thank my family, who always believed in me, for their constant presence in my life and for their unconditional love.

Finally, I would like to dedicate this thesis to Vincenzo who recently passed away for encouraging me to follow my dreams.

Contents

Acronyms

FEP Free Energy Perturbation

FDTI Finite Difference Thermodynamic Integration

GPU Graphic Processing Unit

GPGPU General Purpose Graphic Processing Unit

MC Monte Carlo

MCS Maximum Common Subgraph

MD Molecular Dynamics

MMC Metropolis Monte Carlo

MUE Mean Unsigned Error

NA Non Additivity

PI Predictive Index

R^2 Coefficient of Determination

SAR Structural Activity Relationship

TI Thermodynamic Integration

VdW Van Der Waals

Abstract

The development and approval of new drugs is an expensive process. The total cost for the approval of a new compound is on average 1.0 - 1.2 billion dollars and the entire process lasts about 12 - 15 years [1]. The main difficulties are related to poor pharmacokinetics, lack of efficacy and unwanted side effects. These problems have naturally led to the question if new and alternative methodologies can be developed to find reliable and low cost alternatives to existing practices.

Nowadays, computer-assisted tools are used to support the decision process along the early stages of the drug discovery path leading from the identification of a suitable biomolecular target to the design/optimization of drug-like molecules. This process includes assessments about target druggability, screening of molecular libraries and the optimization of lead compounds where new drug-like molecules able to bind with sufficiently affinity and specificity to a disease-involved protein are designed. Existing computational methods used by the pharmaceutical industry are usually focused on the screening of library compounds such as docking, chemoinformatics and other ligand-based methods to predict and improve binding affinities, but their reliable application requires improvements in accuracy.

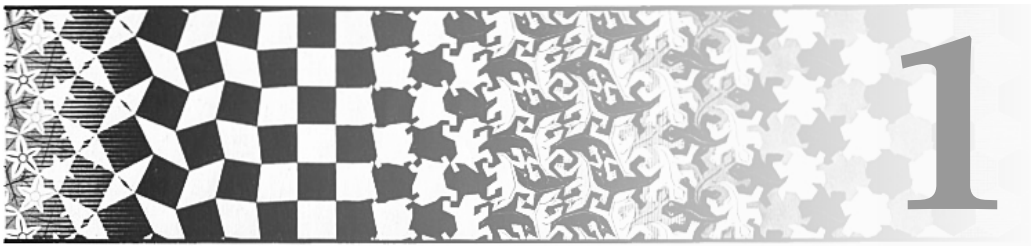
New quantitative methods based on molecular simulations of drug binding to a protein could greatly improve prospects for the reliable *in-silico* design of new potent drug candidates. A common parameter used by medicinal chemists to quantify the affinity between candidate ligands and a target protein is represented by the free energy of binding. However, despite the increased amount of structural information, predicting binding free energy is still a challenge and this technique has found limited use beyond academia. A major reason for limited adoption in the industry is that reliable computer models of drug binding to a protein must reproduce the change in molecular conformations of the drug and protein upon complex formation and this includes the correct modelling of weak non-covalent interactions such as hydrogen bonds, burials of hydrophobic surface areas, Van der

Waals interactions, fixations of molecular degrees of freedom solvation/desolvation of polar groups and different entropy contributions related to the solvent and protein interactions. For several classes of proteins these phenomena are not easy to model and often require extremely computationally intensive simulations.

The main goal of the thesis was to explore efficient ways of computing binding affinities by using molecular simulations. With this aim, novel software to compute relative binding free energies has been developed. The implementation is based on alchemical transformations and it extended a preexisted piece of software Sire, a molecular modeling framework, by using the OpenMM APIs to run fast molecular dynamics simulations on the latest GPGPU technology. This new piece of software has equipped the scientific community with a flexible and fast tool, not only to predict relative binding affinities, but also a starting point to develop new sampling methods for instance hybrid molecular dynamics and Monte Carlo. The implementation has been validated on the prediction of relative hydration free energy of small molecules, showing good agreement with experimental data. In addition, non-additive effects to binding affinities in series of congeneric Thrombin inhibitors were investigated. Although excellent agreement between predicted and experimental relative binding affinities was achieved, it was not possible to accurately predict the non-additivity levels in most of the examined inhibitors, thus suggesting that improved force fields are required to further advance the state-of-the art of the field.

"When I was a child I wanted to heal the cherry trees everytime I thought that red fruits were wounded. I was sure that health had abondoned them along with the snow white flowers they had lost. A dream, it was a dream which didn't last for long, so I swore I would have been a doctor and not for a god nor even as a game but for the cherry trees to bloom again, for the cherry trees to bloom again"

— Fabrizio De André. Un Medico



1 Introduction

1.1 An Introduction to the modern Galilean scientific method

INCE human being have been able to reason they have tried to answer questions about nature and the world they live in. In the beginning ideas and questions were formulated in mythological, religious or philosophical contexts. One common feature was their generality typically trying to explain "whatever is" by presenting absolute truths to ancient and definitive questions such as "What life is", "Why are we living". However, a few centuries later the human curiosity took a different path, human being became interested in more "defined" natural phenomena; general questions were avoided and replaced by well defined investigations. This revolution happened in the Hellenic world (507 BC - 323 BC) where Science started its first steps as a set of logical propositions on natural phenomena (Plato's and Aristotle's Academy 428/427 BC - 348/347 BC). This new philosophy aimed to unveil the essential form of the natural world by

conducting a mere and qualitative observation in many cases. The Hellenic ideas were rediscovered during the medieval age when many scientists set the foundation of the scientific method. Galileo Galilei (1564-1642) played a major role in this endeavour. Actually, Galileo never wrote an essay on the scientific method and he never clarified the relationship between what he called "le sensate esperienze" (the experiment) and "le matematiche dimostrazioni" (the mathematical proofs) but his procedures and methods were introduced in several of Galileo's papers and today they form the core of the scientific method. The different stages of the method can be divided into three main processes: the observation, the theory formalization, and the experimental validation.

- The observation is related to the data collection on a specific natural phenomenon. Data is collected on "measurable quantities" i.e. quantities that can be linked to numbers using a measurement process. This was the great innovation compared to the Hellenic Science; the qualitative observations were replaced by quantitative measurements.

- The theory formalization links the measurable quantities by using a mathematical description. The theory should not just be based on the data related to a specific experiment but it should be "deductive" i.e. the theory should be able to predict and explain future observations. However, the theory does not present final answers or absolute truths, and in case of failures, it is replaced by a broader theory which encompasses the previous like a special case without marking an end to this refinement process.

- The experimental validation is the last step of the method. In this stage an experiment is performed under laboratory conditions many times to check if there is agreement between the data predicted by the theory and the experimental observations.

These stages can be represented by using the diagram in Figure 1.1.

In the scientific method the mathematical description which models the observed natural phenomenon is extremely important. Since its introduction, this

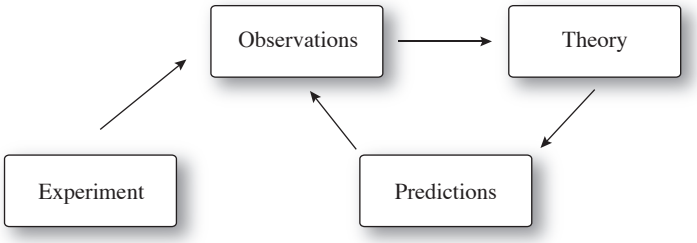


Figure 1.1: *The main stages of the scientific method.*

approach to the rationalisation of the natural world has achieved remarkable successes in many scientific disciplines, but nowadays, the real applicability of the mathematical model to relevant systems is placing significant limits to the theory predictions. In the real world, exact solutions are indeed a notable exception. The major issue is the difficulty to find solutions in a closed form of many problems, which often involve the resolution of partial differential equations e.g. the Maxwell equations, the Schrödinger equation and the N body-problem to cite only few. One of the most famous quotes from Dirac is[2]:

> "The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble"

In order to partially overcome the problem, exact solutions have been replaced by approximations often derived by using numerical methods. In this framework the Turing and von Neumann's ideas on computability[3;4] have found natural growth to the present day, i.e. the application of algorithms to numerically resolve the mathematical model of a theory applied to systems by using computational machines. In other words, in this sense, it is possible to modify the idea itself of scientific method to introdce the concept of simulation, which numerically approximates the theory predictions (Figure 1.2). Currently the main role for simulation is to compare the experimental data with the predicted one; if the level of agreement is not satisfactory the model is judged poor and it needs to be revised. This is a new form of modern experiment where different theories can be

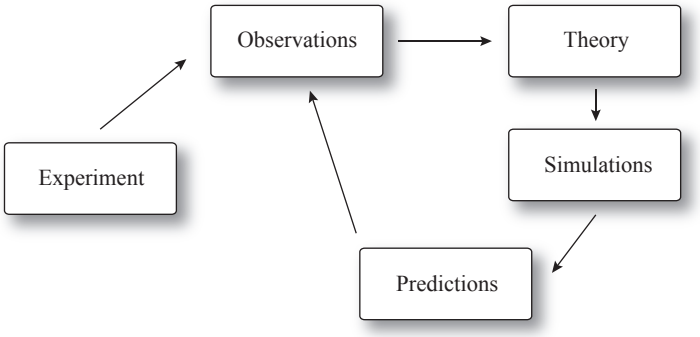


Figure 1.2: *The main stages of the modern scientific method.*

tested before being subjected to a real experiment.

Nowadays, computer simulations are an important tool to complement the traditional approaches to theory and experiment [5]. Nonetheless, simulations can be more that a simple tool: simulations can produce revelations, lead to the unexpected [5], which is extremely important in many contexts to open new research paths. At the beginning of 1970s [5] a heated debate was whether computer simulations have to be considered theories or experiments. The theory side argued that they cannot be considered experiments because no measurement process is involved. On the other side, the experimental position countered that simulation results are used like experiments to check theory validities. In addition, simulation results are prone to problem of reproducibility and statistical errors [5]. How we think about simulations is important. If we accept that simulations are experiments we may be temped to abandon laboratory experiments. The danger lies in severing simulations from reality; it is easy to simulate abstract models whose results have an intrinsic meaning only.

Since their introduction, simulations have permeated many scientific contexts such as Physics, Chemistry and Engineering and more recently Biology. Historically, Metropolis et al. performed the first liquid simulations using the Metropolis Monte Carlo (MMC) method on the Maniac I machine at Los Alamos Laboratory [6]. Nearly at the same time Fermi et al. [7] were able to perform simulations on the anharmonic dynamics of a one-dimensional crystal. Alder performed the first MD simulation on hard sphere dynamics in 1956 [8].

The main theme of this thesis is the application of the newest molecular simulation methodologies to biological systems and in particular the study of free-energy in the protein-ligand binding context. Justifications of the importance of this topic in Drug Discovery are detailed in the following paragraphs.

1.2 Drug Discovery

The drug discovery and development process aims to produce new pharmaceutical drugs to cure illnesses in an effective way. This process is interdisciplinary and, often, involves knowledge from different scientific fields such as Chemistry, Biology, Physics and Computer Science. Due to its importance to human life, many efforts have been made to improve its efficacy and reliability. The idea of drug discovery has its roots in chemotherapy or *therapia sterilisans magna* originating from the beginning of 1900s. The possibility of building "magic bullets", i.e. artificial compounds that were able to kill microorganisms without damaging the host organism, was advanced by P. Ehrlich when he was able to synthesize a compound to cure syphilis[9]. Drug discovery has changed over the last century and, nowadays, the development of new drugs is often performed following a protocol known as the Critical Path[1] (Figure 1.3).

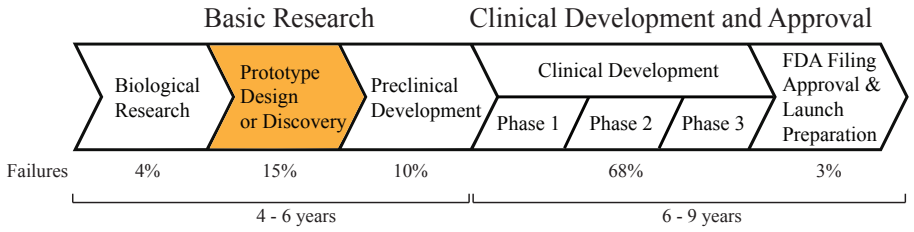


Figure 1.3: *The Critical path of Drug Discovery. A potential drug compound is selected/designed after preliminary biological research on a biomolecule that causes a disease. The drug candidate must complete a series of tests regarding its potential and safety. On average 5000 - 10000 compounds are submitted and evaluated for each candidate that finishes the pathway[1]. The main failures are in the clinical tests and, on average the approval time is between 12-15 years. A key point along the path is the design of new ligands highlighted in orange.*

In the critical path two main different stages can be distinguished: basic re-

search and clinical development/approval. In the basic research stage a biomolecular target is identified; this is often a biological molecule thought to be involved in a specific disease. Potential new drug-like molecules which are able to modulate the biological function of the target are then designed and synthetised. Subsequently tests are performed in vitro and in vivo in the Preclinical Development stage to assess efficacy, pharmacology and toxicity. These studies are usually based on models that are thought to be predictive of the Clinical Development stage where, new candidate compounds are assessed in different phases on human beings and, finally are approved for the market. The whole drug discovery path lasts about 12 - 15 years and the total cost is on average 1.0 - 1.2 billion dollars[10]. It is interesting to observe the percentages of failures along the critical path. Sams et al.[11] analysed the limitations of the target-based drug discovery approach showing the dramatic falling of new drug approval. The main factors for these failures are related to poor pharmacokinetics, lack of efficacy, animal toxicity, side effects on human beings and market issues (Figure 1.3). This has naturally led to the question as to whether this approach is efficacious and more importantly if new and alternative optimisation methods can be developed.

Along the critical path, the design/discovery of new drug-like molecules that are able to bind macromolecule targets activating or inhibiting specific biological functions is an important point and it is highlighted in Figure 1.3. In this stage, the quick and correct design of potent and selective ligands could drastically reduce failures and cut down costs and time and, for this reason, it is the main research area of this study.

The design of new-drug like molecules can be performed using different approaches. In the High Throughput Screening (HTS) a large number of biological modulators and effectors are screened and assayed against selected and specific targets by using different types of libraries including combinatorial chemistry, genomics, protein, and peptide[12]. The main goal is to accelerate drug discovery by screening large compound libraries at a rate that may exceed a few thousand compounds per day or per week and for this reason is frequently used by the pharmaceutical industry[12]. Other popular approaches are the Ligand-Based and the

Structure-Based. In the first case (indirect drug design) a model of the biological target is inferred based on the knowledge of binders to the specific target and this model is used to design new ligands. On the other hand, in the second approach ligands are designed using structural information of the target, usually known from crystallographic or NMR data and, this methodology has also been used in this investigation. In Structure-Based Drug design, promising selected hits undergo more extensive optimisation steps known as the lead-optimisation stage. In it an important aspect is to quantify the affinity between candidate ligands and a target protein by determination of dissociation constant K_d. This constant is linked to the change of the Gibbs free energy through the equation:

$$\Delta G = RT \ln \frac{K_d}{c_0} \ , \tag{1.1}$$

where R is the ideal constant gas, T the temperature and c_0 the standard state concentration. In equation 1.1 the free energy change is evaluated between the thermodynamic state where the biomolecular target P and the drug-like molecule L are in a solvent environment, and the state where the biomolecule and the drug-like molecule form the solvated complex PL. It is possible to represent this dynamic equilibrium using the process:

$$P + L \rightleftharpoons PL \tag{1.2}$$

The optimization of protein-ligand interactions relies most of the time on structural modifications of previously identified promising ligands. In this stage, known as Structural Activity Relationship (SAR), the binding affinity is often optimised by trying to improve its components i.e. enthalpy ΔH and entropy ΔS:

$$\Delta G = \Delta H - T\Delta S \ . \tag{1.3}$$

Many different factors influence enthalpy and entropy. Enthalpy contributions are related to inter- and intra- molecular interactions such as electrostatic or ionic bonds, hydrogen bonds, Van Der Waals (VdW) interactions, dipole-dipole

interactions [13]. Entropy components are connected with the fixations of molecular degrees of freedom and solvation/desolvation of polar groups [13]. The optimisation of all these components is not an easy task. For example VdW forces are optimised by the shape complementarity between the biomolecular target and the drug-like molecule [13], while hydrogen bonds are optimised when hydrogen bond donors and acceptors have optimal geometries in the complex. Entropy changes are related to the reduction of translational, rotational and internal degrees of freedom of ligand and protein in complexation and the release of water molecules from the binding site. Freire et al. [13] have shown how the pharmaceutical industry developed inhibitors of HIV protease from 1996 to 2006 (Figure 1.4). Their analysis showed that the binding affinity of the

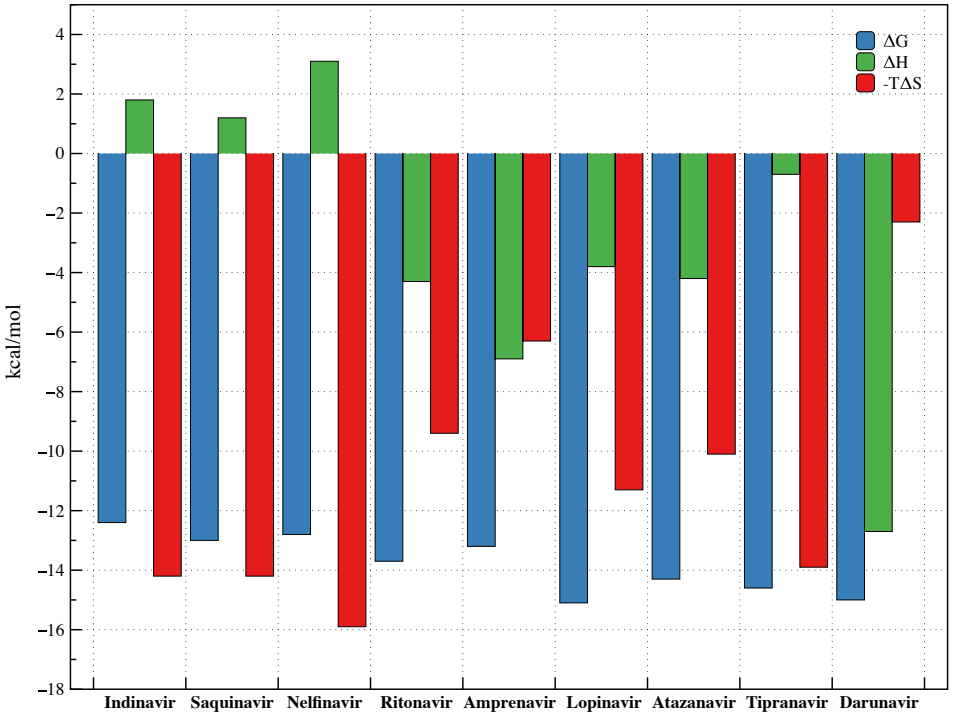


Figure 1.4: *The thermodynamic signature of all HIV-1 protease inhibitors developed and approved by FDA from 1996 to 2006. Although the free energy of binding did not change considerably, the newest inhibitors, which have a better efficacy, are often optimised in entropy, enthalpy or both components. Figure adapted from Freire et al.* [13].

earlier HIV-PR drugs was dominated by entropy, whereas the binding affinity of

the later and, also, more effective had a stronger enthalpic component.

The previous analysis has shown the importance of the entropic and enthalpy components in the design of new drugs. However, a deep understanding of these contributions cannot be achieved without considering the protein-ligand dynamics. Indeed, it is notorious that in some cases ligand binding can produce significant conformational changes to the ternary or quaternary structure of a protein[14], therefore a full understanding of the binding process cannot be achieved without accounting for the dynamic aspect. As a consequence, it is not only important to consider the average structure of a molecule in drug design but also consider where dynamic fluctuations take place, their nature and their scale. This is a quite modern view of the topic, probably related to the dominant X-ray crystallographic methods used in the past to provide structural information on bimolecular systems. Indeed, in these methods a static image of a biomolecule is generated with atoms fixed in space. However, this view was abandoned during the "decade of rigid macromolecules"[14] (1965-1975), in favour of a dynamic picture of the bio-molecular systems. A fundamental insight in the problem has been achieved by atomistic simulations where proteins and molecules are represented by a set of interacting particles according to a specified force field. This description often requires many assumptions and limitations due to complexity of the biological systems, indeed their introduction is necessary for a practical use of the model. Despite these drawbacks and limitations, atomistic simulations have been increasingly and successfully used on biological systems in the last thirty years and are expected to become even more prevalent in the future, with the improvements in numerical methods and, at the same time, the increase in computational power. As a consequence, it is important to quantify the level of precision required for any computational method to significantly affect the lead-optimisation efforts of the pharmaceutical industry in a typical workflow in structured based drug design[15]. Usually in the workflow, a computational chemist selects promising ligands for synthesis from a proposed list of hits that lack sufficient affinity. For example a medicinal chemist team might propose 100 ligands and the computational chemist might need to select 10 for synthesis[15].

It has been showed[16] that even a very modest level of computational accuracy can significantly affect the lead-optimisation stage as illustrated in Figure 1.5. In it, the free energies of binding of more than 84.000 compounds against 30 protein targets were examined[17]. It turns out that the distribution of affinities changes in actual compounds proposed by medicinal chemists is very close to a Gaussian distribution centred at zero. If the computational method used to screen the compounds for the synthesis yields correct affinity prediction with a given level of Gaussian random noise then, it is possible to question what level of noise it is possible to tolerate[15]. Screening a fixed number of compounds with this computational methods will find more potent compounds as the computational noise decreases and, therefore, it would be possible to reduce the number of molecules that need to be effectively synthesised. Assuming that it is possible to screen 10 molecules per week how many molecules must be screened to gain a factor of 10 in affinity after filtering by using the computational method? It turn out that with a 0.5 kcal/mol of noise level the number of screened is reduced by a factor of 8; with 1.0 kcal/mol of noise a factor of 5 and even with 2.0 kcal/mol of noise a factor of 3[15]. Therefore, a computational method that could screen 10-100 molecules per week with even 2.0 kcal/mol of noise would improve the lead-optimisation stage by reducing the synthesis needed in a lead series of a factor of 3[15]. So even relatively small numbers of moderately accurate computational predictions may be able to give significant advantage to the pharmaceutical workflow. However, computational prediction to be useful has to be quick. In order to address this aspect, many computational methods based on algorithmic-hardware solutions have been developed such as parallel computing techniques used in this thesis. In addition, the prediction of the binding affinity is in reality just one consideration in lead optimisation. Improvements must be balanced against other main factors such as solubility, permeability, bio-availability[15] to cite only few, which are important physicochemical parameters highly relevant for drug-like performance. Furthermore, drug-like molecules need to be safe and other parameters relevant to toxicology need to be addressed.

The impact of computational methods in drug development and in particular

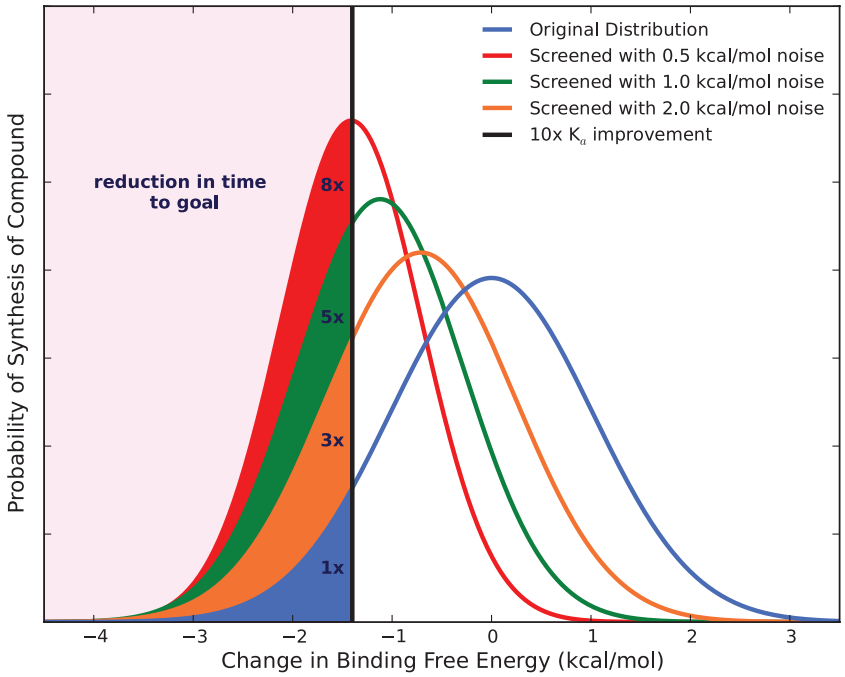


Figure 1.5: *The probability of synthesis of a compound as a function of the binding free energy change for different levels of computational errors. Filled regions indicate those compounds with at least a factor of 10 gain in binding affinity, and are labeled (1x, 3x, 5x, 8x) with the reduction in the number of compounds which would need to be synthesised (on average) to gain this factor of 10 in affinity*[15]*. Blue represents the approximated Gaussian distribution observed experimentally analysing more than 84.000 small compounds against 30 protein targets at Abbot Laboratories*[17]*. Orange, green and red are the distributions generated by filtering the compounds with a hypothetical computational method, which respectively gives correct free energy values with 2.0,1.0, and 0.5 kcal/mol of error noise*[15]*. Even with moderate errors a computational filtering method could drastically improve the efficiency of synthesis in the lead-optimisation stage. Picture adapted from Mobley et al.*[15]

in drug discovery and design has been significant in the last 25 years. Meaningful contributions have been made in lead generation, lead optimisation, prediction of drug likeness, *de novo* design, ligand docking, binding affinity prediction and modulation of ADME (Absorption, Distribution, Metabolism, Excretion) properties and toxicity[18]. Many examples can be found in literature where computational methods played a significant role in drug development. For instance, the ligand-based approach was used to modelling the pharmacophore of benzodiazepine receptor ligands and to design nicotinic agonist using shape matching algorithm[19]. Quantum Mechanics calculations revealed an angiotensin-converting enzyme inhibitor QSAR[19]. Schames et al.[20] by using MD simulations on HIV integrase revealed an unidentified trench that was not evident from available X-ray crystallography and later on was demonstrated that known inhibitors do in fact bind in this cryptic trench[21]. These are just few examples of effective use of computational methods in drug discovery. This thesis has explored relatively new methods in molecular simulations to calculate binding affinities by merging different theoretical methods with the latest hardware technologies and providing new freely available tools to the scientific community. In order to fully understand the approach and the details, the next paragraphs will build up the necessary Statistical Mechanics tools and the computational methods used.

1.3 A brief introduction to Statistical Mechanics

The properties of a macroscopic system which exchanges heat and work with its surrounding are described by the discipline of Classical Thermodynamics. Historically, the development of this science was quite troubled and its general principles were empirically discovered most of the time. A complete and satisfactory theory was only achieved in the 18^{th} century when "heat" was shown to be a form of energy. With the development of modern atomistic theories, a connection was sought between microscopic states of matter and thermodynamic properties. The early attempts tried to apply the classical mechanic method to the atomic description however, this strategy paused difficult problems to overcome. First of all, the relatively high number of degree of freedoms involved in the system de-

scription and the non-linear system interactions between particles. Secondly, the mere use of a classical and "deterministic" description of the atomic dynamics cannot explain many natural facts driven by the second thermodynamic principle. E. Majorana[22] wrote in a broader context:

> "Determinism, which does not leave any rule to human freedom and forces one to consider all the phenomena of life as illusory, implicates a real cause of weakness"

The reconciliation between micro and macro description has eventually been achieved with the description of Statistical Mechanics. A significant understanding was that macroscopic properties could not be strongly dependent on the deterministic dynamic of each single system particle but, rather, on averages through the statistical mechanic notion of "ensemble". The modern view presents a given thermodynamic system characterised as a set of points in the phase space. Each point is known as a microstate and it is modelled with the multidimensional variables:

$$\begin{aligned} \boldsymbol{q} &= (q_1, \cdots, q_n) \ , \\ \boldsymbol{p} &= (p_1, \cdots, p_n) \ , \end{aligned} \tag{1.4}$$

where q_i is the generic generalised coordinate, p_i the related momentum and n the degree of freedom of the considered system. In the phase space, different starting conditions will generate different trajectories. However, many different trajectories in phase space will have the same macroscopic properties. A set of phase space points that share the same macroscopic properties is called an ensemble. If the thermodynamic system visits all its configurations in an infinity amount of time i.e. the system is ergodic then, the physical-chemical thermodynamic observable A can be computed as time averaged in the phase space trajectory as follows:

$$< A >= \lim_{\mathcal{T}\to\infty} \frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} A(\boldsymbol{q}, \boldsymbol{p}) dt \ , \tag{1.5}$$

where the symbol <> denotes the average of the interested property A and $\mathcal{T}$ is the time interval. Usually, dynamic systems are a powerful approach to generate an ensemble and its associated average and they form the basis of the MD

method which is one of the most popular methods to tackle statistical mechanics problems[23]. In the hypothesis of an ergodic system, Gibss[24] suggested that the time average in equation 1.5 can be also computed as average on the ensemble as follows:

$$< A >= \int A(\boldsymbol{q}, \boldsymbol{p})\rho(\boldsymbol{q}, \boldsymbol{p}, t)d\boldsymbol{q}d\boldsymbol{p} \ , \tag{1.6}$$

where $\rho(\boldsymbol{q}, \boldsymbol{p}, t)$ is the ensemble probability density function and, its determination is a key problem in statistical mechanics. Relevant observables are always in thermodynamic equilibrium i.e. variables that do not change in time. As a consequence, the equation 1.6 must lead to a time independent result and this is possible only if $\partial\rho/\partial t = 0$ which will be assumed from now on. In order to completely define a thermodynamic system in equilibrium it is necessary to specify a set of three intensive or extensive properties. Interesting conditions are the NVT (Canonical ensemble), the NPT (Isothermal-isobaric ensemble), the NVE (Microcanonical ensemble) and the μVT (Grand canonical ensemble) where, N is the constant particle number, V the constant system volume, P the constant pressure, T the constant temperature, E the constant energy and μ the constant chemical potential. For each one of these ensembles it is possible to determine the probability density function $\rho(\boldsymbol{q}, \boldsymbol{p})$ and in particular for the NVT ensemble is:

$$\rho_{NVT}(\boldsymbol{q}, \boldsymbol{p}) = \frac{1}{N!h^{3N}} \frac{\exp(-\beta\mathcal{H}(\boldsymbol{q}, \boldsymbol{p}))}{Q_{NVT}} \ , \tag{1.7}$$

where Q_{NVT} is the partition function:

$$Q_{NVT} = \frac{1}{N!h^{3N}} \int \exp(-\beta\mathcal{H}(\boldsymbol{q}, \boldsymbol{p}))d\boldsymbol{q}d\boldsymbol{p} \ . \tag{1.8}$$

In the equations 1.7 and 1.8, $\mathcal{H}(\boldsymbol{q}, \boldsymbol{p})$ is the system Hamiltonian, $\beta = 1/k_B T$ where k_B is the Boltzmann's constant and T the temperature, h the Plank's constant and the term $N!$ is introduced for quantum mechanical reasons to take into account the number of indistinguishable particles in the system. It is interesting to observe that $\mathcal{H}$ cannot be time dependent otherwise the observable A will be time dependent, therefore in this statistical framework are excluded systems where are

present external time dependent forces. If the system potential energy $U(\boldsymbol{q})$ is a function of the generalised coordinate $\boldsymbol{q}$ only, then, the expression of the partition function can be simplified expressing the system Hamiltonian $\mathcal{H}$ as sum of kinetic $K(\boldsymbol{p})$ and potential energy $U(\boldsymbol{q})$ terms. In this case the integral can be decoupled into two independent parts, one dependent on the generalised coordinates and the other part on the momentum coordinates only:

$$Q_{NVT} = \frac{1}{N!h^{3N}} \int \exp(-\beta \boldsymbol{p}^2/2m) d\boldsymbol{p} \int \exp(-\beta U(\boldsymbol{q})) d\boldsymbol{q} \ , \tag{1.9}$$

where m is the mass of each particle. The integral on the momentum part can be solved in a closed form and the simplified version of the partition function is:

$$Q_{NVT} = \frac{1}{N!\Lambda^{3N}} Z_{NVT} \ , \tag{1.10}$$

where $\Lambda = (\beta h^2/2\pi m)^{1/2}$ and Z_{NVT} is the configurational integral:

$$Z_{NVT} = \int \exp(-\beta U(\boldsymbol{q})) d\boldsymbol{q} \ . \tag{1.11}$$

The ensemble average of a pure coordinate-dependent observable $A(\boldsymbol{q})$, can be therefore expressed as:

$$< A >= \frac{\int A(\boldsymbol{q}) \exp(-\beta U(\boldsymbol{q})) d\boldsymbol{q}}{Z_{NVT}} \ . \tag{1.12}$$

From a computational point of view this expression is very attractive, it specifies how to construct the ensemble average of a given thermodynamic observable that can be compared to experimental measurements. An important quantity in Statistical Mechanics is the Helmholtz free energy F linked to the partition function by the equation:

$$F = -k_B T \ln Q_{NVT} \ . \tag{1.13}$$

In terms of free energy component analysis, it is convenient to rewrite the previous equation in terms of ensemble average:

$$\begin{aligned} F &= -k_B T \ln \frac{\int \exp(-\beta \mathcal{H}(\boldsymbol{q},\boldsymbol{p}))d\boldsymbol{q}d\boldsymbol{p}}{\int \exp(+\beta \mathcal{H}(\boldsymbol{q},\boldsymbol{p}))\exp(-\beta \mathcal{H}(\boldsymbol{q},\boldsymbol{p}))d\boldsymbol{q}d\boldsymbol{p}} = \\ &= +k_B T \ln \frac{\int \exp(+\beta \mathcal{H}(\boldsymbol{q},\boldsymbol{p}))\exp(-\beta \mathcal{H}(\boldsymbol{q},\boldsymbol{p}))d\boldsymbol{q}d\boldsymbol{p}}{\int \exp(-\beta \mathcal{H}(\boldsymbol{q},\boldsymbol{p}))d\boldsymbol{q}d\boldsymbol{p}} = \\ &= +k_B T \ln \langle \exp(+\beta \mathcal{H}(\boldsymbol{q},\boldsymbol{p}))\rangle_{NVT} \end{aligned} \tag{1.14}$$

In the context of this study a significant ensemble is also represented by the NPT ensemble. Indeed, chemistry experiment are most of the time conducted at constant pressure, temperature and constant number of particles. In these conditions the expression of the Gibbs free energy is:

$$G = -k_B T \ln Q_{NPT} \ , \tag{1.15}$$

where Q_{NPT} is the partition function of this ensemble:

$$Q_{NPT} = \frac{1}{N! h^{3N}} \int \exp(-\beta(\mathcal{H}(\boldsymbol{q},\boldsymbol{p}) + pV))d\boldsymbol{q}d\boldsymbol{p}dV \ , \tag{1.16}$$

and where p is the pressure and V the system volume.

From now on all the relevant thermodynamic equations useful to this study will be referred and derived in the canonical ensemble, unless otherwise stated, because in it proofs have a convenient notation in this ensemble.

1.4 Mechanical Potential Energy: a very brief introduction

In order to extract interesting macroscopic properties from a system in thermodynamic equilibrium using the ensemble average (equation 1.12), it is necessary to evaluate the system Hamiltonian at different phase space coordinates and, this usually translates in the calculation of the system potential energy $U(\boldsymbol{q})$. Consequently, the specification of the potential energy and its parameters is a key point especially in molecular simulations. It defines the microscopic interaction laws and the agreement with experimental data strongly dependent on its reliability. An accurate description of the interaction laws would require the explicit use of Quantum Mechanics however, the use of this theory is frequently restricted to hundreds of atoms and its application to biological macromolecules is ordinarily impracticable due to the high number of atoms involved in the system description. In the Bohr-Oppenheimer approximation [25] the electron and nuclei motion in a molecule can be separated by splitting the total atomic wave function as product of the electron wave function $\Psi_e(\boldsymbol{r}_i, \boldsymbol{R}_i)$ and the nuclei wave function $\Psi_n(\boldsymbol{R}_i)$ where $\boldsymbol{r}_i$, and $\boldsymbol{R}_i$ are respectively the electron and nuclei positions. This approach relies on the physical fact that the electrons are considerably lighter than the nuclei. The Schrödinger equation for the nuclei is then replaced by Newton's law. The nuclei are then moved according to classical mechanics by using potentials that result from the solution of the Schrödinger equation for the electrons and, many approximations have to be employed [26]. For example, these approximations are derived by using ab-initio methods such as the Hartree-Fock or density functional theory. However, the complexity of the model and the related algorithms enforces limitations on the system size. A further drastic simplification is the use of parametrised analytical potentials that are function of the nuclei positions only. This representation is known as mechanical or atomistic representation. In it the whole atomic system is described as as a charged point mass particle without any internal structure. This approximation suits biomolecular studies where the exact solution of the whole quantum mechanic atom dynamics is numerically impracticable. The functional form of the mechanical potential is usually divided

into two main terms called the "non-bonded" and "bonded" interactions in this context. Three concomitant effects characterised the non-bonded interactions. A repulsive contribution related to the exclusion volume of the VdW forces caused by the inter-atomic repulsions between atomic nuclei. A second contribution describes the dispersion forces or London forces, which is caused by atomic charge fluctuations produced by the presence of another atom creating an attractive dipole-dipole interaction. These two contributions are usually captured by the Lennard-Jones potential U_l:

$$U_l = 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] , \tag{1.17}$$

where r_{ij} is the atomic distance between the atom i and j, ϵ_{ij} and σ_{ij} are respectively the well energy depth and the collision diameter. These parameters are usually determined using experimental and computational approaches such as viscosity data, scattering data and quantum mechanic calculations [14]. Often these parameters are refined also using crystallographic and liquid structures [14].

The final contribution to the non-bonded interactions is related to the electrostatic between particle atoms. These interactions are modelled by using the Coulomb potential providing partial charges to each atom in the system. The potential energy expression U_C is:

$$U_c = \frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{r_{ij}} , \tag{1.18}$$

where q_i is the atom charge of atom i and ϵ_0 the vacuum dielectric constant. The sign and the magnitude of the Coulomb parameters may be obtained by determining the electronic density function of the system from quantum mechanic calculations of the ground state by using first principle methods such as Mulliken [27] and AM1-BCC [28]. Usually the results are basis set dependent and they are improved with experimental data fitting [14]. Coulomb forces are notoriously stronger compared to the Lennard-Jones interactions and, as a consequence, they have a longer-range and are responsible for very relevant hydrogen bond forma-

tions in biomolecular simulations. The intra non-bonded interactions are not usually applied on atoms that are separated by less than three (known as 1-2, 1-3 terms) or four bond lengths. A reasonable explanation is related to the overlapping between the non-bonded interactions and the bonded terms used to delineate the covalent bonding[14]. In molecular simulation, the non-bonded term calculations are expensive with a complexity of $O(N^2)$ where N is the number of particles in the system. In order to reduce their computational cost these are frequently estimated by only considering the atoms in a given cut-off distance only. This approximation is usually improved adding long range correction terms to the Coulomb expression such as the reaction field[29], or using a non cut-off scheme such as the Ewald summation[30]. In addition, in order to avoid discontinuities in energy and/or forces at the cut-off distance, switching functions are regularly used[31].

The bonded contributions to the potential energy are related to vibrational bond terms and energetic rotational terms around bonds. The vibrational terms model the bond and angular stretching dynamics and they are usually accounted by a simple harmonic potential or other anharmonic terms. In biomolecular system, this approximation is motived by constant fluctuations around equilibrium positions at conventional temperatures. The harmonic potential form of the bond U_b and angular U_a vibrational terms are:

$$U_b + U_a = \frac{c_b}{2}(r_{ij} - r_0)^2 + \frac{c_a}{2}(\theta_{klm} - \theta_0)^2 \ , \tag{1.19}$$

where c_b and c_a are respectively the bond and angle force constants while, r_0 and θ_0 are respectively the equilibrium bond distance and the equilibrium angle. Furthermore, the terms r_{ij} and θ_{klm} are respectively the atomic distance between the generic atoms i and j and the angle between the generic atoms k, l and m.

The rotation around bond contributions are modelled by using energetic rotational barriers along bonds. These potential terms U_d are represented by using

a cosine expansion truncated to a lower order:

$$U_d = \sum_n A_n(1 + \cos(n\phi_{ijkl} - \phi_0)) \ , \tag{1.20}$$

where A_n is the height of the torsional barrier, n the multiplicity, ϕ_{ijkl} the dihedral angle formed by the atoms i, j, k and l and ϕ_0 is the equilibrium dihedral angle. This functional form correctly describes two fold barriers such as in amide group and three fold barriers such as in hydrocarbon chains[14]. However for more complex molecule such as sugars, it is necessary to consider other potential expressions[14]. The potential bonded parameters may be determined using quantum mechanics calculations and experimental data. For example the force constant determinations can be performed calculating the minimal structural configuration and then performing a normal vibrational analysis and fitting the results with experimental data[14].

The classical mechanical potential expression in a system can therefore be written as a sum of bonded and non-bonded contributions as follows:

$$\begin{aligned} U &= U_b + U_a + U_d + U_l + U_c = \\ &= \sum_{bonds} c_b(r_{ij} - r_0)^2 + \sum_{angles} c_a(\theta_{ij} - \theta_0)^2 + \\ &+ \sum_{dihedrals} \sum_n A_n(1 + \cos(n\phi_{ijkl} - \phi_0) + \\ &+ \sum_{pairs(i<j)} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{r_{ij}} \ , \end{aligned} \tag{1.21}$$

and excluding from the non bonded terms the 1-2 and 1-3 interactions. Nowadays, different force fields mainly based on the presented force field have been developed, usually for different purposes. The most popular are MM2[32] and its extensions, AMBER[33], CHARMM[34], GROMOS[35] and OPLS[36]. However, many significant physical phenomena are completely ignored in this framework such as polarizability, π-stacking interactions or creation and destructions of chemical bonds. To address these limitations, mixed quantum mechanical-molecular mechanical force fields are under development in a number of laboratories and first

results seem to be encouraging. The thesis author would like to rewrite a quote from the Halgren's paper[37] which, nowadays to some extent is still valid:

> "Some day, consensus on the form and manner of parameterisation of molecular force fields may exist, but for now much remains to be learned. This is as it should be, for the problem being addressed is a hard one: to capture faithfully in a computationally tractable model enough of the real, quantum-mechanical physics to insure that a bio-molecular simulation, properly carried out, will yield a correct answer to useful precision. The way would be clearer if we knew how much physics 'enough' is"

1.5 An introduction to the force field parametrisation

As we have seen, in the context of molecular simulations an important aspect is the force field used to model the molecular system. Usually, force fields are specified by using a potential energy function and a set of parameterisation constants. In the previous paragraph a particular form of the mechanical potential and its parameters was examined but, it is not unique and force fields can differ in the potential energy functional form and the parameters. Usually, the functional form tries to mediate between accuracy and computational efficiency and it is often chosen for its mathematical properties such as the derivability up to the second order, to enable use of MD simulations and minimisation techniques. Generally, force fields are designed to model and predict specific properties for specific classes of molecules e.g. structural or thermodynamic properties and, they could in principle fail to predict other properties when they are applied to molecular systems that differ from a training set[31]. This aspect is know as transferability of the force field parametrisation and it is an important property in the prediction process. The force field parameterisation is usually performed as follows[38]:

- a set of reference data related to chemical and physical properties relevant for the force field application target are selected. These properties have to be computable and, in many cases, these are computed by using coordinates

and particle momenta. Examples of reference data are[38]: chemical structures determined by using X-ray crystallography, microwave spectroscopy and electron diffraction; electrostatic properties such as dipole moments; physical properties such as density; thermodynamic quantities such as enthalpy of formation, hydration free energy and heat capacities;

- divide the data in training and validation set;
- set a starting value for each parameter to be optimised accordingly to their chemical and physical meaning in the training set. During the optimisation a check on the parameter values is required to ensure that the parameters are still meaningful;
- refine the parameter values until an optimal parameterisation is reached. Usually in this process, the minimisation of a selected scoring function is involved. A popular scoring function $\mathcal{F}(\mathbf{p})$ is[38]:

$$\mathcal{F}(\mathbf{p}) = \sum_{l}^{N_0} w_l (O_l^{calc}(\mathbf{p}) - O_l^{ref}(\mathbf{p}))^2 \ , \tag{1.22}$$

  where $\mathbf{p}$ is the vector parameter, N_0 the number of reference data points, w_l are the reference data weights, $O_l^{calc}(\mathbf{p})$ and $O_l^{ref}(\mathbf{p})$ are respectively the calculated data and the reference data value itself;
- the optimised parameters are then validated on the validation set to predict the reference data. This step enables evaluation of the force field transferability.

The parameterisation of a force field by using all possible available reference data and the potential energy function parameters is computationally unfeasible. For instance, if the considered potential energy is the mechanical potential energy 1.21 and N is the number of atoms in a system then, the total number of unique bond equilibrium distances and bond force constants to optimise is $\binom{N+2-1}{2}$ each. The same number is required for the optimisation of each LJ parameters ϵ and σ. For the atomic charges N optimised parameters are required and, $\binom{N+3-1}{3}$

are required for each angle equilibrium distance and angle amplitude. Finally, $\binom{N+4-1}{4}$ optimised parameters are required for each dihedral angle amplitude and phase. If the previous analysis is restricted to the first $N = 100$ elements of the periodic table[39], then the total number of required parameters is $\geqslant 10^9$, which is hardly feasible to obtain. In order to make parametrisation tractable, it is possible to select a subset of significant atoms present in many organic molecules such as: H, C, N, O, F, Si, P, Cl, Br and I[39]. However, results are often not satisfactory because it is necessary to take into account the local chemical environments to derive accurate force fields[31]. Thus, many force fields rely on concept of "atom type". For example, in the MM2[32], MM3[40] and MM4[41] force fields, eight carbon atomic types are distinguished: sp^3, sp^2, sp, carbonyl, cyclopropane, radical, cyclopropene and carbonium ion[31]. In two widely used bio-molecular force fields such as OPLS[36] and AMBER[33], 41 atomic types are defined[39] and in the UFF[42] force field, which includes most of the elements of the periodic table, 126 atom types are described[39]. Another frequently used strategy to reduce the parametrisation effort is to exclude some parameters from the optimisation process. For instance, the atomic charges can be computed by using quantum mechanics calculation methods prior to runnig a MD simulation. A disadvantage of this approach is that atomic partial charges are no longer transferable and must be computed for every new molecule to model. Finally, another interesting method is the approach where parameters that depend on more than one atom are re-defined as function of a single per atom parameter. For example, in the LJ potential the parameter σ_{AB} is function of the two atoms A and B. However, by using the mixing rule:

$$\sigma_{AB} = \frac{\sigma_A + \sigma_B}{2} \ , \tag{1.23}$$

where σ_A and σ_B are single per atom parameters, the total number of optimised parameters required is N instead of the original $N(N+1)/2$.

1.6 Free Energy Calculations

The rigorous theory of free energy calculation was developed many years ago and due to the limited computational resources at that time numerical applications to the theory were very limited[43]. The roots of the early method can be ascribed to John Kirkwood[44] when in his work on the derivation of the integral equation in liquid state theory he introduced the the notion of order parameter to infer free energy difference between two thermodynamic states. Nearly 20 years later, Zwanzig[45] showed how to calculate free energy changes by using ensemble averages, which forms the theoretical basis of the popular Free Energy Perturbation (FEP) method. The initial applications of free energy calculations had to wait to more computational power to be extended to physical and chemical relevant systems and therefore, the calculations were originally domain of analytical studies[43]. A remarkable non trivial success of these starting attempts was the explanation of the hydrophobic effect in the work of Pratt and Chandler[46] which was subsequently confirmed by numerical simulations. When computational power became available, a plethora of free energy calculations applied to increasingly complex molecular systems began to flourish[43]. Most of the initial efforts were based on MMC approaches e.g. initial applications to Lennard-Jones fluids[47], study of atomic clusters[48] and ion hydration investigations[49]. In 1979 two studies faced the nature of the hydrophobic effect by using free energy calculations. Susumu Okazaki et al.[50] used MMC to estimate free energy of hydrophobic hydration. They found that the hydrophobic hydration was accompanied by a decrease in internal energy and large entropy loss in agreement with the conventional picture of the phenomenon[43]. In the second, study Bruce Berne et al.[51] investigated a multistage approach applied to model system formed by Lennar-Jones spheres in a bath of water molecules and they successfully recovered the results of Pratt and Chandler[46] related to the hydrophobic interactions[43]. In the 1980s new research directions were explored. Tembe and McCammon[52] applied the FEP method to model ligand-receptor assemblies. Jorgensen and Ravimohan[53] estimated the hydration free energy of ethane and methanol by

applying a common topology shared between the transformation end points. Another remarkable example of calculation[43] was the study of the S_N2 chemical reaction of $Cl^- + CH_3Cl$ in gas phase and solution by Chandrasekhar et al.[54], which laid out the basis of modern QM/MM calculations. In the same year Kollman et al. opened new paths to in-silico modelling of site-directed mutagenesis by using the FEP approach to calculate free energy changes associated with point mutations in amino acid side chains[55;56;57;43]. They employed a first attempt of slow-grow method, which was rigorously theoretical formulated 10 years later by Jarzynski[58]. It is also worth to mention the approached developed by Fleischmand and Brooks[59] to calculate entropy and enthalpy differences[43]. They showed that the error associated with these two thermodynamic quantities in calculations were about one order of magnitude higher that the corresponding free energy error. In contrast to FEP the Thermodynamic Integration (TI) method was used only in the late 1980s when Straatsma and Berensden[60] successfully used the method to calculate ionic hydration by mutating neon to sodium and, nowadays, this technique is one of the most common free energy approaches.

The initial studies reported very good agreement between experimental and predicted free energy data however, after the starting enthusiasm it was soon realised that these successes were probably due to good fortune rather than precise computer simulations[43]. For example, in many cases it was observed that predicted free energy deviates from the related experimental value as soon as more sampling was accumulated. In addition, many systems appeared to be non-ergodic showing slow-convergence issues. These and other observations have mainly driven the different research paths in the last 20 years and many successes were achieved relaxing the "theoretical rigour"[43] by using "well-motivated"[43] approximations. For example, in many stratification free energy calculation strategies, the path connecting the end points of the simulation was divided into sub-paths and, often, numerical instabilities occured at the reference and target states. Beutler et al.[61] introduced the concept of soft core potential to mitigate the problem and, nowadays, it is frequently used in alchemical free energy calculations where creation and annihilation of chemical groups are involved. Another common problem

was the free energy dependence on the system size in cases where significant electrostatic interactions were present[43]. The use of reaction field[29] and Particle Mesh Ewald lattice[30] considerably mitigated the problem in neutral systems. On the other hand, in charged systems, Hummer et al.[62] showed that the system-size dependence can be faced if a self-interaction term is taken into account in the simulation. This term is associated with the interaction of charged particles with their periodic images. Hummer in his paper was able to correctly calculate the hydration free energy of a sodium ion in a water box consisting of 16 water molecules only. Another problem that caused concern in the field was the use of holonomic constraints. In numerical calculations these are often used to remove high frequency vibrations allowing the equation of motions to be integrated by using larger time steps. In the early years of free energy calculations the effect of frozen internal degree of freedom was ignored[63] however, it was showed that constraints could significant alter the accessible volume phase space and, therefore, influence free energy simulations[43]. Stefan Boresch and Martin Karplus[64] showed the importance of the metric tensor corrections that can be analytically estimated in many systems.

Practical applications of free-energy calculations to the pharmaceutical context were limited due to the computational cost and accuracy. The primary line of research in this direction targeted drug design applications. Eric Duffy and W. Jorgensen[65] simulated a set of 200 pharmaceutical organic compounds in aqueous environment calculating their solvation free energy. With the increase of computational power W. Jorgensen used the FEP method in the lead-optimisation stage to design new potent anti-HIV-1 agents[66]. A compromise between accuracy and high throughput was proposed by David Perlman and Paul Charifson[67] suggesting that one step FEP simulation on a grid surround the ligand gives a roughly estimate of the binding constant. A significant boost to reach throughput was also achieved with the introduction of reliable implicit solvent models. Simonson et al.[68] showed how to approximate long range interactions in a continuum solvent environment without sacrificing accuracy, which led to a significant reduction of the computational cost in atomistic simulations. More

recently Andrew McCammon et al.[69] employed the Poisson-Boltzmann surface area (MM/PBSA), successfully facing the problem of estimating conformational changes in free energy upon binding of a ligand to its receptor[43]. An aspect in free energy calculation that causes considerably difficulty has been the system sampling. In the 1990s different approaches based on the treatment of an order parameter as dynamic variable were developed[43]. The idea was to construct a series of MD trajectories or MC walks with a different value of the order parameter. The probability of visiting the different system states characterised by a different value of the order parameter can be significantly different. Occasionally, a configuration swap is attempted between the systems, accepting or rejecting the swapping based on a Metropolis criterion. A suitable parameter used to increase the smoothing of the probability distribution function is the temperature and, this method known as parallel tempering has became increasingly popular to tackle difficulty problems where high-energy barriers are present between the different system states[43]. Another technique was proposed by Laio and Parrinello[70] in 2002. They developed the metadynamics approach based on the definition of collective variables to efficiently explore the free energy surface. In it, a memory kernel guaranties that the different visited free energy minima in the free energy landscape are progressively filled as the simulation progresses in the long run[43].

As previously seen one of the most concerning problem in free energy calculations performed on biomolecular systems with a high number of degree of freedoms is the necessity to efficiently search the phase space. One common characteristic of such systems is the presence of energy barriers lower or higher than the simulated system thermal energy[43]. In an ergodic system every point in the phase space have to be accessible from every other point. This requirement for complex systems simulated along conventional time scales could produce disconnected phase space regions and, therefore, systems are often trapped in "un-escapable states". Hodel et al.[71] showed how errors produced by insufficient sampling could impact free energy calculations[43]. It turn out than, even for relatively small systems such a nine-residue peptide improper sampling resulted in errors of the order of 1 kcal/mol, which accounted for about 50% of the total calculated free energy[43].

Many methods have been designed to enhance the sampling to yield more accurate results than conventional approaches such as MD and MMC. However, the development of new algorithms, often based on new mathematical approaches is just one possible way to tackle the problem. It is believed[43] that this strategy lead to non "perfect" methods but, rather methods that perform better for particular applications of problems[43]. Nonetheless, a simplest approach to enhance the sampling exists and it significantly emerge from the previous brief free energy history i.e. the use of brute force methods based on the enhancement of computational power. In the last 25-30 years many progresses have been made in the computational field e.g. the processor performance exponential growth. This growth is the result of two main factors: increase in the processor complexity related to higher device density (number of transistor per chip) and the introduction of new architectural features such as large cache memories, large instruction buffers, multiple instructions per cycle, multi-threading, branch predictions to cite only few[72]. However, nowadays, there are physical limits that have be reached in the computational development such as the finite speed of signal propagation along a wire and heat dissipation issues. One possible path settled to tackle these problems by the semiconductor industry has been the use of multiple processors.

Since 2003, the semiconductor industry has followed two main directions to designing microprocessors: the multicores and the many-cores[73]. The multicores began as two-core CPU processors, with the number of cores approximately doubling with each semiconductor process generation[73]. A current exemplar is the recent Intel Core i7 microprocessor, which has four processor cores, each of which support multiple-instructions, hyperthreading and is designed to maximise the execution speed of sequential programs[73]. In contrast, the many-core architecture focuses more on the execution throughput of parallel applications and a current exemplar is the GPU GeForce GTX Titan X with more than 3000 cores, each of which is a heavily multithreaded[73]. The development of this technology has been driven by the market demand for high-quality, real-time graphics computer applications such as video games and animated movies and ad-hoc microprocessors named Graphic Processing Unit (GPU) were developed to this end. There

is a large computational gap between CPUs and GPUs in particular in numerical applications. The design of a CPU is made by sophisticated control logic units to deal with the operating system requests. On the other hand, the design philosophy of the GPUs is shaped by the video game industry, which exerts tremendous economic pressure for the ability to perform a massive number of floating-point calculations per video frame. In addition, graphics chips have been operating at approximately 10 times the memory bandwidth of contemporaneously available CPU chips[73]. GPUs are designed as numeric computing engines and therefore they have been adopted in many scientific fields where to validate scientific hypotheses TFLOPS performance is often required and PFLOPS performance would be highly desirable[72].

As previously seen the sampling in biomolecule systems is one of the main concerning problems. This thesis has explored and applied the computational power of the modern GPU processors to face the sampling problem in biomolecular simulations. We have seen that it is very useful to quantify the relative binding affinities between prominent ligands in the lead-optimization stage. With this aim a fast relative free energy implementation based on alchemical transformations and the FDTI method has been developed. The implementation merged two existing piece of software Sire[74] and OpenMM[75] gaining flexibility from the advanced Sire molecular modeling framework and speed from the OpenMM APIs performing molecular dynamics simulations directly on the modern GPUs. The produced implementation is the starting point to perform new science merging flexibility and computational power in a simple, efficient and effective way. This is often missing in many competitor software restricted to predefined and rigid schemes, which do not fit in a flexible frame so important in the scientific research context.

1.7 FEP and TI

As described in the previous paragraphs, the binding free energy is a relevant property to improve the drug efficacy at the early steps of drug discovery. Molecular simulations are nowadays used to support rational drug design but, the prediction of free energy of binding and its components, still remains the "Holy Grail" of Computational Chemistry [76]. In the last thirty years, many computational methods have been developed to compute binding affinities, usually balancing a trade-off between accuracy and computational cost. In order to fully understand the difficulties related to free energy calculation, it is necessary to examine how free energies are computed. The evaluation of the absolute Helmholtz free energy involves the numerical calculation of equation 1.13, which is rewritten here for clarity:

$$F = -k_B T \ln Q \ .$$

This equation is extremely important. It describes the link between macroscopic and microscopic worlds evaluating the macroscopic observable F througth the calculation of the partition function Q, which is involved the microscopic description. The effective determination of Q is unworkable for realistic cases, because it requires the calculation of the system accessible phase space volume integral, for example by using molecular simulations. On the other hand, in molecular systems we are frequently interested in the calculation of free energy differences between two thermodynamic states A and B. In this case from equation 1.13 it follows:

$$\Delta F = -k_B T \ln \frac{Q_B}{Q_A} \ , \tag{1.24}$$

which by using the simplification described in § 1.3 can be rewritten as ratio of configurational partition functions:

$$\Delta F = -k_B T \ln \frac{Z_B}{Z_A} \ . \tag{1.25}$$

In general the two thermodynamic states A and B can differ in many ways, such as extensive and intensive variables e.g. temperature, pressure or volume, or

they can be modelled by two different Hamiltonians $\mathcal{H}_A$ and $\mathcal{H}_B$ e.g. a residue mutation in a protein or functional groups in a ligand. The determination of the ratio Q_B/Q_A can be simplified as follows:

$$\begin{aligned}
\Delta F &= -k_B T \ln \frac{\int \exp(-\beta \mathcal{H}_\mathcal{B}) d\boldsymbol{q} d\boldsymbol{p}}{\int \exp(-\beta \mathcal{H}_\mathcal{A}) d\boldsymbol{q} d\boldsymbol{p}} = \\
&= -k_B T \ln \frac{\int \exp(-\beta \mathcal{H}_\mathcal{B}) \exp(\beta \mathcal{H}_A) \exp(-\beta \mathcal{H}_A) d\boldsymbol{q} d\boldsymbol{p}}{\int \exp(-\beta \mathcal{H}_\mathcal{A}) d\boldsymbol{q} d\boldsymbol{p}} = \\
&= -k_B T \ln \frac{\int \exp(-\beta (\mathcal{H}_B - \mathcal{H}_A)) \exp(-\beta \mathcal{H}_A) d\boldsymbol{q} d\boldsymbol{p}}{\int \exp(-\beta \mathcal{H}_\mathcal{A}) d\boldsymbol{q} d\boldsymbol{p}} = \\
&= -k_B T \ln \int \exp(-\beta (\mathcal{H}_B - \mathcal{H}_A)) \rho_A d\boldsymbol{q} d\boldsymbol{p} = \\
&= -k_B T \ln < \exp(-\beta (\mathcal{H}_B - \mathcal{H}_A)) >_A
\end{aligned} \tag{1.26}$$

In the previous proof, it is assumed that the temperature between the two states A and B is the same, but it is straightforward to generalise this condition. The expression of ρ_A in the previous equation is the probability to find the system A in the microstate $(\boldsymbol{q}, \boldsymbol{p})$ in the NVT ensemble. Due to its importance, equation 1.26 is named the free energy perturbation formula[44;45] and it allows the calculation of ΔF monitoring the factor $\exp(-(\mathcal{H}_B - \mathcal{H}_A)/k_B T)$ e.g. by using molecular simulations. This methodology is known as FEP because the Hamiltonian of the system B can be written as a perturbation of the Hamiltonian of the system A:

$$\mathcal{H}_B = \mathcal{H}_A + \Delta \mathcal{H} \ , \tag{1.27}$$

where $\Delta \mathcal{H}$ is the perturbed energy term. However, this approach could present a substantial issue i.e. the configurational phase spaces related to the thermodynamic states A and B could not truly overlap. This effectively means, that the configurations generated during the sampling of the state A, for example using molecular simulations could not be significant configurations of the state B and the Hamiltonian change in the equation 1.26 could be large, producing a low contribution to the exponential term and, therefore, a possible poor convergence in the calculation of ΔF. In general, this issue can be tackled by dividing the thermodynamic path A $\rightarrow$ B in closer intermediate steps where the Hamiltonian changes are expected to be small and the whole free energy change between the

end states is computed as sum along the path steps.

Another popular method used to compute free energy change is the TI method. In this approach a thermodynamic system in the state A is transformed into the thermodynamic system B by changing a coupling parameter λ defined in a given range $[\lambda_A, \lambda_B]$ where, λ_A and λ_B respectively represent the system in the states A and B. The coupling parameter can control the change of extensive and intensive properties or can mutate the Hamiltonian of system A into B ($\mathcal{H}_A \rightarrow \mathcal{H}_B$), or it could control structural changes between A and B such as values of torsional angle. If the Hamiltonian depends on a coupling parameter then, the free energy change can be written as:

$$\Delta F = \int_{\lambda_A}^{\lambda_B} \frac{\partial F}{\partial \lambda} d\lambda \ , \tag{1.28}$$

and differentiating the equations 1.13 respect to λ[77]

$$\frac{\partial F}{\partial \lambda} = \int \frac{\partial \mathcal{H}}{\partial \lambda} \rho \ d\boldsymbol{q} d\boldsymbol{p} = < \frac{\partial \mathcal{H}}{\partial \lambda} >_\lambda \ , \tag{1.29}$$

and therefore,

$$\Delta F = \int_{\lambda_A}^{\lambda_B} \frac{\partial F}{\partial \lambda} d\lambda = \int_{\lambda_A}^{\lambda_B} < \frac{\partial \mathcal{H}}{\partial \lambda} > d\lambda \ . \tag{1.30}$$

In the TI method the calculation of ΔF is evaluated using the ensemble average $< \partial \mathcal{H}/\partial \lambda >$ in contrast with the FEP approach where is evaluated the ensemble average $< \exp(-\beta \Delta \mathcal{H}) >$. The TI method should not suffer of the FEP problem related to the poor sampling of the final state and, as a consequence, it is expected to produce better convergence. However, the TI method also requires the integral calculation and usually this is numerically performed by using quadrate rules or polynomial regression and, these techniques could bias the method. In addition, the numerical integral accuracy depends on the shape of $< \partial \mathcal{H}/\partial \lambda >$ producing acceptable results if it is smooth enough. In particular transformations, known as Alchemical transformations, the LJ and Coulomb interactions can be abruptly turned on or off, simulating the appearance or disappearance of particles in a system and, therefore, the change $< \partial \mathcal{H}/\partial \lambda >$ could be significant producing poor numerical integration. Due to their importance in this study, alchemical

transformations and the TI methods are further discussed in Chaper two where will be presented an implementation of the TI free energy method.

1.8 Sampling methods: MD and MC

As previously seen, the determination of macroscopic thermodynamic properties using microscopic quantities relies on the calculation of the ensemble averages evaluated on different phase space configurations, usually generate by using Molecular Dynamics or Monte Carlo sampling methods. In these methodologies the simulation frequently begins by selecting a set of initial conditions and then the system is evolved to produce new phase space configurations. Generally, the first simulation segment corresponds to an equilibration stage and, at the point where the equilibrium is numerically achieved (steady fluctuations), the previous history is discharged. Subsequently, significant physical properties are calculated from the ensemble average and, finally, the simulation is allowed to proceed until there is no significant variation among the investigated physical properties. The main characteristics of MD and MC are briefly detailed below.

1.8.1 Molecular Dynamics

In an atomistic simulation, atoms are identified as a set of N interacting particles in a given force field and, in this case, the system can be modelled by using the Cauchy's problem related to Newton's second order linear differential equations:

$$\begin{aligned} \boldsymbol{F}_i &= m_i \frac{d^2 \boldsymbol{r}_i}{dt^2} \ , \\ \boldsymbol{r}_i(t_0) &= \boldsymbol{r}_i^0 \ , \\ \left.\frac{d\boldsymbol{r}_i}{dt}\right|_{t_0} &= \boldsymbol{v}_i^0 \end{aligned} \tag{1.31}$$

$\forall i = 1, \cdots, N$. In the previous system $\boldsymbol{F}_i$ is the generic force acting on particle i, m_i its mass, $\boldsymbol{r}_i$ the particle position as function of the time parameter t, $\boldsymbol{r}_i^0$ and $\boldsymbol{v}_i^0$ are respectively the particle position and velocity at the starting time t_0. In the MD method the force acting on each particle can be derived from the potential

energy function as follows:

$$\nabla_{\boldsymbol{r}_i} U = -\boldsymbol{F}_i \ . \tag{1.32}$$

Using the previous equation it is possible to prove that the total system energy is constant and, therefore, the natural ensemble suitable to MD is the microcanonical ensemble NVE. Equations 1.31 are known as the N-body problem and, in the general case, it is not possible to find a solution in a closed form. Numerical methods are then used to find approximate solutions. A very naïve method to integrate the system 1.31 is to evaluate the acceleration $\boldsymbol{a}(t)$ of each particle by using the potential $U(\boldsymbol{r}(t))$ and, subsequently, it is possible to update the velocities $\boldsymbol{v}(t + \Delta t)$ and the positions $\boldsymbol{r}(t + \Delta t)$ supposing that the acceleration is constant between t and $t + \Delta t$ where Δt is the time step. The procedure is then reiterated. This very simple algorithm suffers of many issues producing poor sampling without conserving the total system energy. More suitable integrators are the Verlet schemes i.e. Verlet-Störmer [78] and velocity Verlet [79]. MD can also be performed in other frequently used ensembles. In the NVT case, it is necessary to control the temperature in the system. This can be achieved in many ways and, a very simple approach is offered by the velocity scaling techniques [80]. The temperature is indeed related to the kinetic energy [31]:

$$\sum_i \frac{1}{2} m_i v_i^2(t) = \frac{3}{2} N k_b T(t) \ . \tag{1.33}$$

In order to control the temperature at a reference temperature T_{ref}, it is possible to multiple the velocities by a scaling factor λ:

$$\sum_i \frac{1}{2} m_i (\lambda v_i(t))^2 = \frac{3}{2} N k_b T_{ref} \ . \tag{1.34}$$

Therefore, the difference in temperatures obtained subtracting the two previous equations can be written as:

$$\Delta T = (\lambda^2 - 1) T(t) \ . \tag{1.35}$$

Multiplying at each time step the velocities by the factor

$$\lambda = (T_{ref}/T_{curr})^{1/2} \ , \tag{1.36}$$

it is possible to keep the temperature constant at T_{ref} while T_{curr} is the temperature at the selected step calculated by using equation 1.33. Another scaling technique is to couple the system to a thermal bath [81]. In this case the rate of change of the system temperature is modelled using the equation [31]:

$$\frac{dT}{dt} \sim \frac{\Delta T}{\Delta t} = \frac{1}{\tau}(T_{bath} - T(t)) \ , \tag{1.37}$$

where τ is the parameter that describe how tightly the system is coupled to the bath and T_{bath} is the bath temperature. Therefore, ΔT in this case is:

$$\Delta T = \frac{\Delta t}{\tau}(T_{bath} - T(t)) \ , \tag{1.38}$$

and the scaling factor λ using the equation 1.35 is:

$$\lambda^2 = 1 + \frac{\Delta t}{\tau}\left(\frac{T_{bath}}{T(t)} - 1\right) \ . \tag{1.39}$$

However, the previous two scaling approaches are not able to correctly sample the NVT ensemble producing poor averages. Other approaches have been developed based on stochastic processes [82] or extended system methods [83] and analogue methodologies have been developed to control the pressure.

1.8.2 Monte Carlo

The effectively determination of the equation 1.12 requires the calculation of multidimensional integrals that, for practical system cannot be evaluated using standard techniques e.g. quadrature rules. For example the Trapezium and Simpson's rules to integrate an n-dimensional function f on a given volume require to evaluate m^{3n} points where, m is the number of selected approximation points in each dimension [31]. Hence, if we just select a 50-dimensional function evaluated with 3 points per dimension this would require $3^{150} \sim 10^{71}$ calculations, which is cur-

rently unachievable by any computational machine [31]. The Monte Carlo method and its variations have been developed to calculate multidimensional integrals. The basic idea is to generate many random points in the multidimensional space and count the number of points that fall inside the integral volume. In particular this technique applied to the equation 1.12 leads to the following algorithm [31]:

1. randomly generate 3N space coordinates;
2. calculate the potential energy function $U_i(\boldsymbol{q})$ at step i on the 3N coordinates and then evaluate $\exp(-\beta U_i(\boldsymbol{q}))$;
3. accumulate the previous values and return to step 1.
4. After N_{iter} iterations the ensemble average of an observable A is estimated as follows:

$$< A >= \frac{\sum_i^{N_{iter}} A_i(\boldsymbol{q}) \exp(-\beta U_i(\boldsymbol{q}))}{\sum_i^{N_{iter}} \exp(-\beta U_i(\boldsymbol{q}))} \ . \tag{1.40}$$

However, this approach is not very efficient. Indeed, most of time random arrangements of atoms will have a negligible Boltzmann factor $\exp(-\beta U_i(\boldsymbol{q}))$. As a consequence, the low-energy states, which weight more in the ensemble average sum, are not efficiently sampled. The Metropolis Monte Carlo approach [6] is a significant improvement to the original Monte Carlo method. In this technique, the low-energy states are efficiently explored (importance sampling) by using Markov Chains. In the canonical ensemble, the algorithm evolves molecular systems from a state where the particles have positions $\mathbf{q}$ to a new randomly generated position $\mathbf{q}'$ using an acceptance probability $\mathcal{A}(\mathbf{q}'|\mathbf{q})$ described by the Metropolis criterion [23]:

$$\mathcal{A}(\mathbf{q}'|\mathbf{q}) = \min\left\{1, e^{-\beta(U(\mathbf{q}')-U(\mathbf{q}))}\right\}. \tag{1.41}$$

The acceptance criteria is based only on the variation in the potential energy involved in the specific move. If the potential energy change is positive the move is accepted with probability one otherwise, it is accepted with a probability that exponentially decreases with the increase in the potential energy change [23]. However, the displacement of all the particles in a single move could lead to a low

acceptance rate, and this problem becomes more severe as soon as the particle number increases[23]. An immediate solution is to move one particle at a time in a move randomly selected among all the particles. This strategy also has a positive effect on the potential energy calculation. It is only necessary to update the potential energy with the contribution related to the new particle position in the case of a positive acceptance test.

1.9 Chapter summary and thesis overview

This chapter introduced the main theme of this thesis, which is related to modern and new computational methods used to help and speed up the drug discovery process. A brief overview of the novel drug development process was detailed and in particular the importance of the drug discovery and design process was highlighted to cut down cost and time along the critical path. The importance of the binding affinity as one of the key parameter used by medicinal chemists to improve drug efficacy was described and how computational predictions of this quantity could significantly help the synthesis of new potent drugs. An introduction of relevant thermodynamic quantities was explained to connect the micro atomic world with macroscopic observables through the use of the ensemble average notion. Finally, computational methods to predict free energy change were introduced. The FEP and TI approaches were explained together with molecular simulation techniques such as MD and MMC.

The following chapter will present an implementation of the single topology and the FDTI methods used in conjunction to calculate relative free energy changes. Applications of the implemented code will be used to compute relative and absolute hydration free energy by using the single and the dual topology methods. Chapter three will detail the importance of the flexibility in a simple molecular system and how the force field parameterisation can impact on the molecular flexibility; a comparison with experimental data will be detailed. In Chapter four the non-additivity phenomenon in drug design will be investigated. The relative free energy implementation will be used to predict the relative binding affinities of series of congeneric Thrombin inhibitors and to estimate the non-additivity level

present in the systems. The final chapter will try to explain the non-additivity origins selected Thrombin inhibitors. Finally, conclusions will summarised the whole thesis work.

"Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective positions of the beings which compose it, if moreover this intelligence were vast enough to submit these data to analysis, it would embrace in the same formula both the movements of the largest bodies in the universe and those of the lightest atoms; to it nothing would be uncertain, and the future as the past would be present to its eyes"

— Pierre-Simon Laplace. Introduction to Oeuvres vol. VII, Théorie Analytique de Probabilités

2 Free energy calculations using Alchemical Transformations

2.1 An Introduction to Alchemical Transformations

IN the first chapter the FEP and TI methods were introduced to calculate free energy changes between two selected thermodynamic states and the notion of coupling parameter λ was introduced. This parameter is used to transform a given thermodynamic system between two end states e.g. changing the intra- and inter molecular interactions, structural modifications or even extensive or intensive parameters. In general, in the free energy context, transformations where the coupling parameter is used to convert a thermodynamic system between different chemical states are known as alchemical transformations. This is in part due to the fact that these transformations could also involve changes in molecular species and therefore, in a sense, this is the realisation of the inacces-

sible dream of the proverbial alchemist to transmute matter[43] chased during the middle age. Historically alchemical transformations are grounded in the works of Kirkwood[84;44] and Zwanzig[45]. The former introduced the coupling parameter technique, and described how to use it in conjunction with the TI method to calculate the free energy change. Later on, Zwanzig showed how to calculate changes in the free energy by evaluating exponentials of potential energy differences over an ensemble average of system configurations. From a computational point of view the coupling parameter is introduced in the system Hamiltonian $\mathcal{H}$ and the Hamiltonian shape is changed between the selected end states along the alchemical transformation. The coupling parameter is usually selected in a given range $[\lambda_A, \lambda_B]$ with the constraint that $\mathcal{H}(\lambda_A)$ and $\mathcal{H}(\lambda_B)$ respectively describe the starting and the final thermodynamic states involved in the transformation. The choice of the lambda dependence in the system Hamiltonian, and therefore, the alchemical path between the end states is arbitrary. However the transition could be computationally very different among the paths. Indeed, the selection of a path with high-energy barriers between the starting and final states could lead to a very inefficient calculation. Figure 2.1 (a-c) illustrates some specific alchemical mutations and Figure 2.1 (d) elucidates the concept of thermodynamic paths between two end states.

In the alchemical paradigm another important aspect is the notion of thermodynamic cycle. Free energy is indeed a thermodynamic state function and its total variation along a closed thermodynamic path must equate zero ($\Delta F_{A \to A} = 0$). This fact is frequently used to evaluate free energy changes between thermodynamic states separated by high-energy barriers. Indeed, in principle, it is possible to introduce new thermodynamic states along a closed cycle where it is easier to calculate free energy changes and then, indirectly calculate the free energy change between the original states (Figure 2.2). Thermodynamic cycles are reversible, and therefore, forward and backward transformations can be considered. The thermodynamic cycle closures are frequently used as quality check for free energy calculations and the discrepancy from zero measures the transformation hysteresis[43]. If the hysteresis is greater than the statistical uncertainties then

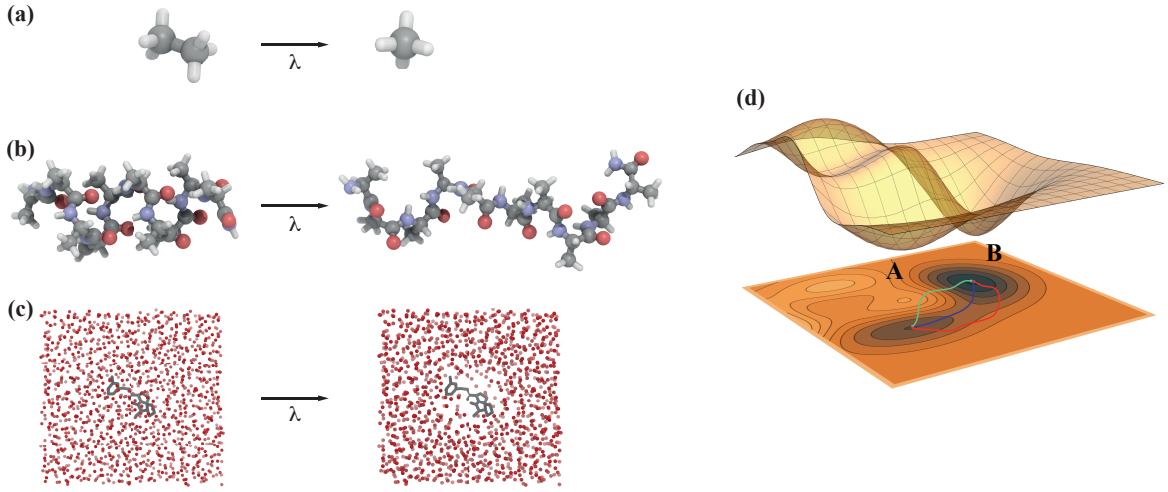


Figure 2.1: *(a) The coupling parameter λ can be used to mutate a molecule into another molecule. In this case an ethane molecule is mutated into a methane molecule. (b) The coupling parameter can also be used to control structural modification. In this case a deca-alanine molecule is forced to unfold by changing torsion angles during an "alchemical" mutation. (c) λ can also be used to change the interactions between molecules. In this case the inter-molecular interactions between a solute and a solvent are progressively switched on from a non-interacting to a full-interacting case. This technique can be used to compute a hydration free energy. (d) The free energy surface of a thermodynamic system can be represented as a hyper dimensional surface. In this case the free energy is function of two variables only and two minima A and B are highlighted. The minimum in B is a stable state while A is a meta-stable state. The system can switch between the two states using different thermodynamic paths. Although the free energy difference is the same between the two states, the path used to switch between A and B could present different computational demand and therefore can be less or more efficient.*

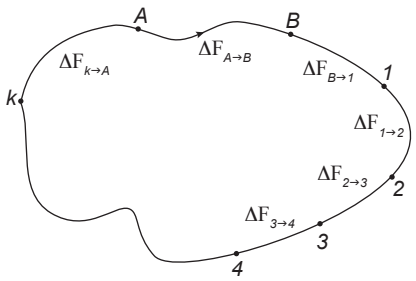


Figure 2.2: *Free energy changes along a closed thermodynamic path must be zero. Therefore, the free energy change between two thermodynamic states* A *and* B *can be indirectly evaluated as follows:* $\Delta F_{A\to B} = -(\Delta F_{B\to 1} + \Delta F_{1\to 2} + \cdots + \Delta F_{k\to A})$.

a potential ergodic issue could be present along the calculation. In addition, low hysteresis and low statistical uncertainties do not guarantee that the transformation is correct; the alchemical transformation could be still biased by incorrect force field parameterization[43].

An example of alchemical transformation, which will be further investigated in this chapter is represented by the alchemical transformation used to compute the hydration free energy. This important quantity measures the free energy change required to hydrate a molecule from the gas phase to an aqueous environment. Figure 2.3 (a) represents the starting and final thermodynamic states usually needed to calculate the hydration free energy (Gibbs Free energy). In the first thermodynamic state the molecule is in the gas phase while in the final thermodynamic state the molecule is fully interacting with the solvent. Along this path the intermolecular interactions between the solute and solvent are progressively switched on. An alternative path used to compute the hydration free energy is also presented in Figure 2.3 (b). In this case the hydration free energy is evaluated as difference between two annihilation free energy changes. The first is the free energy change required to switch-off the solute intramolecular interactions in the gas phase while the latter is the energy change required to switch-off the solute intra- and the solute-solvent inter-molecular interactions.

The absolute free energy of binding between two molecules can be calculated in a very similar fashion. In the protein-ligand binding context the binding affinity

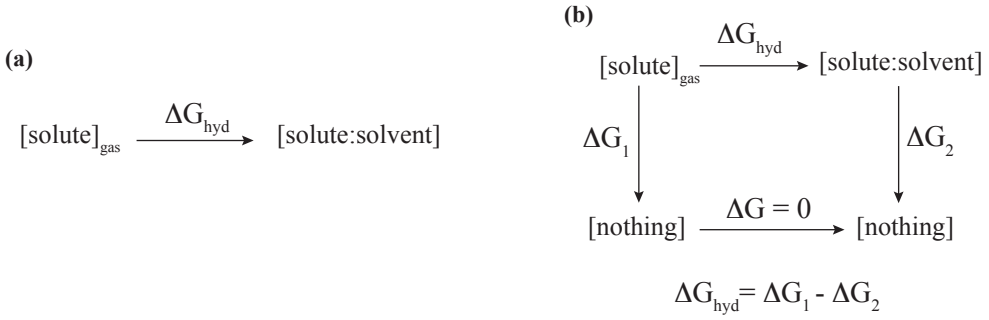


Figure 2.3: *Different thermodynamic paths can be used to compute free energy changes. (a) The hydration free energy of a molecule can be calculated progressively switching on the intermolecular interactions between a solute and a solvent. (b) The hydration free energy is calculated using a double annihilation in this case.*

is the free energy change required to form a complex in an aqueous environment starting from an unbound state Figure 2.4 (a). The direct calculation of the absolute binding affinity is particularly difficult by using the path in Figure 2.4 (a). Effectively, it is necessary to simulate a binding event where the ligand and the protein meet during their diffusive motions in a solvent environment and then start the binding process; overall this requires the dehydration of the binding site and configurational changes of the protein and the ligand and the whole process could be very difficult to simulate in-silico. An alternative path

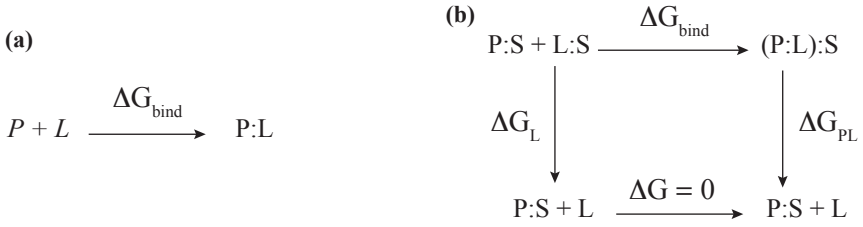


Figure 2.4: *(a) Absolute binding affinity can be computed simulating a direct biding event. However, this is very computationally demanding and alternative paths are usually used. (b) The binding affinity can be calculated performing a ligand (L) de-solvation and decoupling the ligand from the solvent (S) and the protein (P) binding site environment. The ligand decoupling could potentially be difficult and distance harmonic restraints are usually required.*

used to calculate the absolute binding affinity between a ligand and receptor is shown in Figure 2.4 (b). In this case the binding affinity is calculate as free energy difference required to de-solvate the ligand and to decouple the ligand from the solvent and the receptor. It interesting to observe that in both final states where the free energy change is zero (Figure 2.4 (b)) the intra-molecular ligand interactions are not annihilated. The selected pathway could also present some issues. Indeed, the ligand decoupling from the protein could potentially lead the ligand to stick in different protein parts when close to its full decoupling state producing difficult free energy convergence. In order to overcome this and other issues, it is useful to constraint the orientational configuration of the ligand related to the receptor applying distance harmonic constraints in the decoupling and de-solvation transformations and analytical corrections are required to the free energy calculation to take into account these contributions [85;86].

Absolute binding affinity is an important property as discussed in the first chapter, but more often medicinal chemists are interested in the relative binding affinity between ligands. Indeed, there is often the need to compare free energy of bindings between two ligands L_1 and L_2. The problem can be represented through the thermodynamic processes:

$$\mathrm{P} + \mathrm{L}_1 \rightleftharpoons^{\Delta G_1} \mathrm{PL}_1$$
$$\mathrm{P} + \mathrm{L}_2 \rightleftharpoons^{\Delta G_2} \mathrm{PL}_2$$

The free energy change $\Delta\Delta G = \Delta G_2 - \Delta G_1$ is the relative free energy of binding. The thermodynamic cycle in Figure 2.5 is frequently used to compute relative binding affinity between two ligands L_1 and L_2 with a given host protein P. The relative binding affinity is evaluated as difference between two alchemical transformations. In the first the ligand L_1 is mutated in the ligand L_2 in a solvent environment (ΔG_L) while, in the latter, the ligand L_1 is mutated in the ligand L_2 in the binding site (ΔG_{PL}). Because of their importance in rational drug design this thesis mainly focused on relative binding affinity calculations.

In order to perform alchemical transformations, generally two different method-

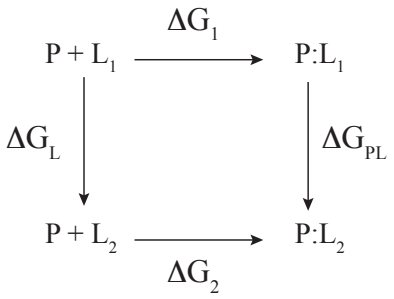


Figure 2.5: *The thermodynamic cycle used to calculate the relative free energy of binding between two ligands L_1 and L_2. In this case $\Delta\Delta G = \Delta G_2 - \Delta G_1 = \Delta G_{PL} - \Delta G_L$*

ologies are implemented: the dual and single topology. In the dual topology[87] method a system where two complete versions of two changing molecules respect to the coupling parameter λ coexist at all the time. One version of the molecule represents the initial state and the other the final end point. The atoms of the two molecule versions interact with the rest of the system in an appropriate weighted-mix however, the topologies do not interact each other. In the method the atom types and internal coordinates of the coexisting molecules never change[87]. The coupling parameter λ is usually introduced in the potential energy function of the system Hamiltonian $\mathcal{U}(\mathbf{r})$ and represents the scaling constant defining the topology mixtures at any intermediate point. Usually the weighted-mix is defined as follows[23]:

$$\mathcal{U}(\mathbf{r}, \lambda) = f(\lambda)\mathcal{U}_A(\mathbf{r}) + g(\lambda)\mathcal{U}_B(\mathbf{r}), \tag{2.1}$$

where $\mathcal{U}_A$ and $\mathcal{U}_B$ are the potential energies of the end states A and B, and the functions f and g satisfy the constraint: $f(\lambda_B) = g(\lambda_A) = 0$ and $f(\lambda_A) = g(\lambda_B) = 1$[23]. Frequently λ_A and λ_B are respectively selected as 0 and 1 and in this thesis it will be assumed from now on otherwise explicitly stated. The two functions f and g are also completely arbitrary but they are often selected as linear: $f = (1 - \lambda)$ and $g = \lambda$.

In contrast with the dual topology approach, in the single topology paradigm the change is formulated in terms of a system where the atom types and target

internal coordinates are modified to reflect the end states. The transformation is practically implemented changing per-atoms, bonds, angles and dihedral force field parameters controlled by using the coupling parameter. A shared topology is used between the alchemical end points and, usually the larger topology involved in the mutation is used as common scaffold for the end states. The eventually omitted structural atoms between the starting and final end states are treated as vanishing particles and are often referred as "dummy particles".

It has been proved that in alchemical mutations where bond lengths are changed as a part of a free energy calculation a contribution from those changes could be necessary to include by using the so called PMF contribution [63]. Bond changes are frequently involved in the single topology method and it could be necessary to take into account these corrections. On the other hand, in the dual topology method the target bond lengths in the hybrid analytic potential function do not change with λ and therefore free energy calculations could be simplified in this schema. However, the dual topology approach could require greater conformational rearrangement of the system at every increment in λ, since a greater number of atoms interact with the system compared to the single topology method and therefore, the single topology method could result in less abrupt changes at the endpoints and possibly lead to faster convergence.

Pearlman [87] compared both the approaches in free energy calculation by using FEP and TI for different alchemical transformations. The study clearly indicates that the free energy simulations performed by using the single topology approach converge more quickly than those using dual topologies. The difference is particularly acute when relatively short calculations are carried out or when the FEP method is used [87]. Comparing TI and FEP for the same system, it appears that the two methods are approximately comparable with the single topology model but that TI is appreciably more efficient with the dual topology system [87]. In addition, there are cases where the dual topology method is preferable e.g. in systems where closed ring changes (purine and pyrimidine nucleic acid bases or where an aromatic to aliphatic substitution is performed for a protein side chain). In such cases the slower rate of the dual topology method must be adjusted and

the sampling performed accordingly.

As previously described, in alchemical transformations, atoms can frequently appear or disappear and often this could lead to the so called end point catastrophe issue. Indeed, the inter-atomic interactions between atoms that vanish and the surrounding environment could become extremely intense due to possible steric clashes generated by the short interatomic distances. This fact could potentially lead to severe numerical instabilities during the simulations which it is necessary to correctly handle in a practical implementation of the single and dual topology method. The next paragraphs will detail the implementation of a relative free energy calculation code by using the FDTI method and the single topology paradigm.

2.2 The FDTI method

The FEP and TI methods were briefly introduced to calculate free energy changes between two thermodynamic states in Chapter one. In particular the TI method describes how free energy differences can be evaluated by calculating the ensemble average through the equation 1.30 and, re-written here, as Gibbs free energy:

$$\Delta G = \int_{\lambda_A}^{\lambda_B} < \frac{\partial \mathcal{H}}{\partial \lambda} > d\lambda \ . \tag{2.2}$$

Due to the high number of degrees of freedom the previous equation cannot be analytically resolved in realistic systems and numerical techniques are often applied. Equation 2.2 can also be rewritten as follows:

$$\Delta G \equiv \int_{\lambda_A}^{\lambda_B} \frac{\partial G}{\partial \lambda} d\lambda \ . \tag{2.3}$$

In the method equation 2.3 is numerically approached by opportunely selecting a set of N values of the coupling parameter in a range $[\lambda_A, \lambda_B]$ and the integral in the equation 2.3 is approximated by using quadrature rules or polynomial regression. For example the integral can be approximated by using the simple trapezium

formula as follows:

$$\Delta G = \int_{\lambda_A}^{\lambda_B} \frac{\partial G}{\partial \lambda} d\lambda \simeq \frac{\lambda_B - \lambda_A}{2N} \sum_{k=1}^{N} \left(\frac{\partial G}{\partial \lambda}\Big|_{\lambda_{k-1}} + \frac{\partial G}{\partial \lambda}\Big|_{\lambda_k} \right) . \tag{2.4}$$

In order to perform the numerical integration, it is generally required the evaluation of the integrand function on a set of selected values of the coupling parameter, i.e. it is necessary to calculate:

$$\frac{\partial G}{\partial \lambda}\Big|_{\lambda_k} . \tag{2.5}$$

In the FDTI method these values are numerically estimated by using the finite difference central derivate formula:

$$\frac{\partial G}{\partial \lambda} \simeq \frac{\Delta G(\lambda \to \lambda + \Delta\lambda) - \Delta G(\lambda \to \lambda - \Delta\lambda)}{2\Delta\lambda} , \tag{2.6}$$

where $\Delta\lambda$ is a selected parameter, which controls the precision of the gradient calculation. In the literature[88] it is possible to find better approximations for the gradient but the computational cost is usually higher and the improvements are often negligible. In order to compute the free energy differences in the equation 2.6 the Zwanzig's formula[45] is used in the FDTI method:

$$\Delta G(\lambda \to \lambda \pm \Delta\lambda) = -1/\beta \ln\langle \exp[-\beta(\mathcal{U}(\lambda \pm \Delta\lambda) - \mathcal{U}(\lambda))]\rangle_\lambda , \tag{2.7}$$

where the symbol <> represents the ensemble average generated by using MMC or MD. Algorithm 1 and Figure 2.6 represent an implementation of the FDTI method.

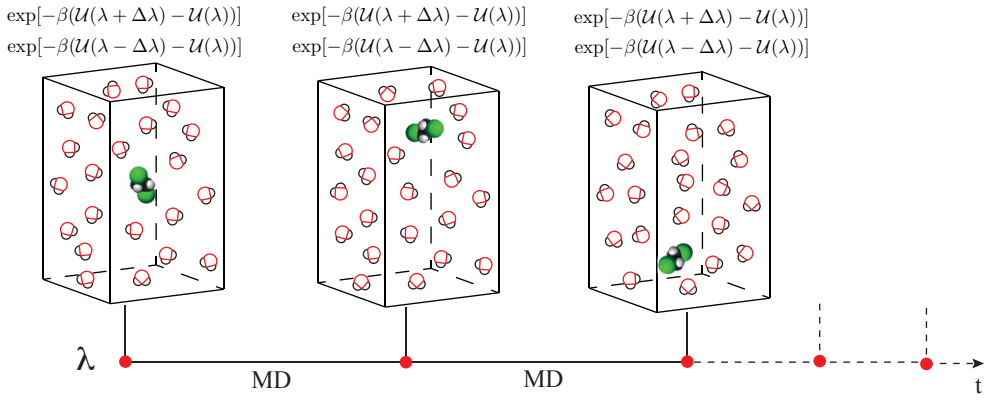


Figure 2.6: *In order to calculate free energy change between two thermodynamic states the algorithm 1 can be used. An ensemble of configurations for a selected value of the coupling parameter* λ *can be generated by using MD. At beginning the system potential energy is calculated on the starting configuration. Subsequently, the value of* λ *is perturbed in* $\lambda \pm d\lambda$*. The system potential energy function is then calculated on these values* $\mathcal{U}(\lambda \pm d\lambda)$ *and the exponential differences can be computed* $\exp[\mathcal{U}(\lambda \pm d\lambda) - \mathcal{U}(\lambda)]$ *and stored. The value of the coupling parameter is then set to its original value and the system in evolved in time by using MD for a selected number of steps. The new positions of the atoms in the space will define new values for the potential energy* $\mathcal{U}(\lambda)$ *and its variations* $\mathcal{U}(\lambda \pm d\lambda)$ *and, therefore, new exponential differences. The procedure can be iterated and all the exponential values can be used to compute the ensemble average and, as a consequence, the value of the free energy gradient at the selected coupling parameter value.*

Algorithm 1 The FDTI algorithm

1: Select a set of λ values: $\{\lambda_A = \lambda_1, \lambda_2, \cdots, \lambda_n = \lambda_B\}, \lambda_k \in [\lambda_A, \lambda_B]$
2: $\lambda \leftarrow \lambda_1$
3: **for** $i \neq n$ **do**
4: Generate an ensemble of configurations $\{\mathcal{C}_1, \mathcal{C}_2, \cdots \mathcal{C}_l\}_\lambda$ for the selected λ
5: **for** $j \neq l$ **do**
6: Evaluate and save the exponentials: $\exp[-\beta(\mathcal{U}(\lambda \pm \Delta\lambda) - \mathcal{U}(\lambda))]$
7: **end for**
8: Compute the ensemble average: $-1/\beta \ln\langle\exp[-\beta(\mathcal{U}(\lambda \pm \Delta\lambda) - \mathcal{U}(\lambda))]\rangle_\lambda$
9: Calculate the free energy change: $\Delta G(\lambda \rightarrow \lambda \pm \Delta\lambda)$
10: Calculate $\left.\frac{\partial G}{\partial \lambda}\right|_\lambda$
11: $\lambda \leftarrow \lambda_{i+1}$
12: **Go to** step 3:
13: **end for**
14: Numerically compute the integral by using $\left.\frac{\partial G}{\partial \lambda}\right|_{\lambda_k}$ values

2.3 An Implementation of the Single Topology Method

The FDTI method can be used in conjunction with the single topology method to calculate relative binding free energy. Two alchemical mutations (Figure 2.5) are usually required to calculate relative binding affinity between two ligands L_1 and L_2. However, it is necessary to explain what "mutation" means in this context and how a molecule can be transformed into another molecule. Following the Born-Oppenheimer approximation[25], atoms can be represented as point particles interacting with each other in a given force field described by using an appropriate potential energy function. In principle, an atom species can be mutated into another atom species and, therefore, a molecule can be mutated into another molecule, if the intra- and inter- molecular interactions between the two species are changed between the end points of two thermodynamic states. In the single topology method this is achieved by introducing the coupling parameter lambda

in the total mechanical potential energy function (equation 1.21) as follows:

$$\begin{aligned} U(\lambda) &= U_b(\lambda) + U_a(\lambda) + U_d(\lambda) + U_l(\lambda) + U_c(\lambda) = \\ &= \sum_{bonds} c_b(\lambda)(r_{ij} - r_b(\lambda))^2 + \sum_{angles} c_a(\lambda)(\theta_{ij} - \theta_a(\lambda))^2 + \\ &+ \sum_{dihedrals} \sum_{n} A_n(\lambda)(1 + \cos(n\phi_{ijkl} - \phi_d(\lambda)) + \\ &+ \sum_{pairs(i<j)} 4\epsilon_{ij}(\lambda) \left[\left(\frac{\sigma_{ij}(\lambda)}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}(\lambda)}{r_{ij}} \right)^{6} \right] + \frac{1}{4\pi\epsilon_0} \frac{q_i(\lambda) q_j(\lambda)}{r_{ij}} \end{aligned} \tag{2.8}$$

where each force field parameter is explained in Chapter one §1.4. The functional form of the force field parameters respect to the coupling parameter λ can be arbitrary and in the implemented code was selected as linear:

$$\begin{aligned} c_b(\lambda) &= c_b^1\lambda + c_b^0(1-\lambda) \ , \\ r_b(\lambda) &= r_b^1\lambda + r_b^0(1-\lambda) \ , \\ c_a(\lambda) &= c_a^1\lambda + c_a^0(1-\lambda) \ , \\ \theta_a(\lambda) &= \theta_a^1\lambda + \theta_a^0(1-\lambda) \ , \\ A_n(\lambda) &= A_n^1\lambda + A_n^0(1-\lambda) \ , \\ \phi_d(\lambda) &= \phi_d^1\lambda + \phi_d^0(1-\lambda) \ , \\ \epsilon_{ij}(\lambda) &= \epsilon_{ij}^1\lambda + \epsilon_{ij}^0(1-\lambda) \ , \\ \sigma_{ij}(\lambda) &= \sigma_{ij}^1\lambda + \sigma_{ij}^0(1-\lambda) \ , \\ q_i(\lambda) &= q_i^1\lambda + q_i^0(1-\lambda) \ . \end{aligned} \tag{2.9}$$

The superscript indexes "0" and "1" respectively denote the force field parameters in the starting and final states.

In an alchemical mutation between two molecules it is possible to distinguish three categories of atoms. If L_1 and L_2 are respectively the starting and the final molecule it is possible to have:

- atoms that are in L_1 but not in L_2. This category of atoms is called "to dummy" atoms;
- atoms that are in L_2 but not in L_1. This category of atoms is called "from

dummy" atoms;

- atoms that are both present in L_1 and L_2. This category of atom is called "hard" atoms.

During the mutation in the "to dummy" atom group the intra- and inter-molecular interactions are progressively turned off between these atoms and the rest of the system. In the final state these atoms are "to dummy" or "ghost" particles, they do not present any charges and Lennard-Jones parameters. On the other hand, in the "from dummy" atom group the intra- and inter- molecular interactions are progressively turned on. The "hard" atom group describes cases where atoms do not change between the end states or if they change they mutate into different atom types. During the mutation for this group the intra- and inter- atomic interactions are progressively changed between the starting and the final atom configurations. Figure 2.7 shows the three atom groups considering alchemical mutations between the ethane and methane molecules.

The "to dummy" and "from dummy" atom groups are introduced to deal with cases where particles appear or disappear along an alchemical transformation. From a computational point of view these atoms can lead to numerical instability if they are not correctly handled. In particular near the end points of the mutations the interatomic distances between "to dummy" or "from dummy" atoms and other system atoms e.g. solvent molecules could become relatively small (steric clashes) producing high variation in the inter-molecular potential energy and therefore possible computational failures. In order to mitigate the problem the use of the soft core potential is advisable. A convenient functional form of the this potential is[89]:

$$U_{soft}(\lambda) = \sum_{pairs(i<j)} 4\epsilon_{ij}(\lambda)\left[\frac{\sigma_{ij}^{12}(\lambda)}{(\lambda\delta\sigma_{ij}(\lambda)+r_{ij}^2)^6} - \frac{\sigma_{ij}^{6}(\lambda)}{(\lambda\delta\sigma_{ij}(\lambda)+r_{ij}^2)^3}\right] + \\ + \frac{(1-\lambda)^n q_i(\lambda)q_j(\lambda)}{4\pi\epsilon_0}\frac{1}{\sqrt{\lambda+r_{ij}^2}} \tag{2.10}$$

where σ_{ij} and ϵ_{ij} are the Lennard-Jones parameters of a pair of particles i and j whose distance is r_{ij} while q_i, q_j are the atomic charges; δ and the integer

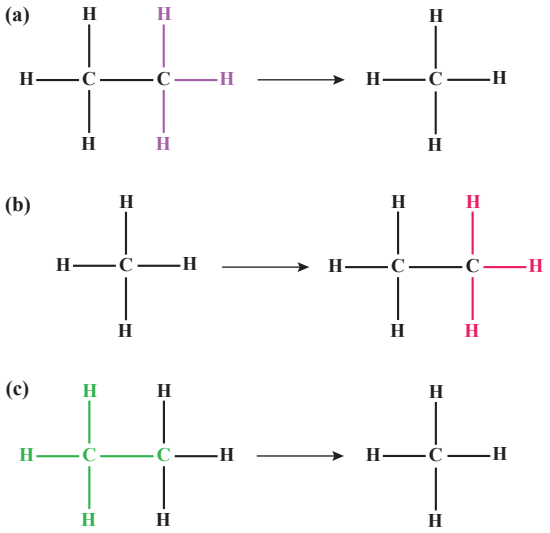


Figure 2.7: *(a) Mutation of ethane to methane. The purple atoms are "to dummy" atoms because they are present in the starting molecule but not in the final one. (b) Mutation of methane to ethane. The red atoms are "from dummy" atoms because they are present in the final molecule but not in the starting one. (c) Same mutation as in (a). The green atoms are "hard" atoms because they are present in the starting and in the final molecules. In this case one of the two carbons is mutated into a hydrogen atom.*

n are used to control the softening. From this equation it is clear that when the distance between atoms is zero (steric clashes) the potential is numerically computable (finite) because of the presence of the term $\lambda\delta\sigma_{ij}$. It is interesting to observe that the coupling parameter λ needs to vary differently between the three groups of atoms. For example in the "to dummy" group, for increasing values of coupling parameter the interactions are progressively turned off in contrast with the "from dummy" group where they are progressively turned on.

In the implemented code, the coulomb interactions were also evaluated by using the reaction field potential[90;91]. The use of this potential simulates the presence of a medium of constant dielectric $\epsilon_{solvent}$ to partially correct the Coulomb long-range interactions computed with cut-offs. Although this method is not as accurate compared to other methods such as the Ewald summation[30] it is very easy to implement and very computationally efficient. The Coulomb interactions in the equations 2.8 and 2.10 can be replaced by the equations:

$$U_{crf}(\lambda) = \sum_{pairs(i<j)} \frac{q_i(\lambda)q_j(\lambda)}{4\pi\epsilon_0}\left(\frac{1}{r_{ij}} + k_{rf}r_{ij}^2 - c_{rf}\right) \tag{2.11}$$

$$U_{soft-crf}(\lambda) = \sum_{pairs(i<j)} \frac{(1-\lambda)^n q_i(\lambda)q_j(\lambda)}{4\pi\epsilon_0}\left(\frac{1}{\sqrt{\lambda + r_{ij}^2}} + k_{rf}(\lambda + r_{ij}^2) - c_{rf}\right) , \tag{2.12}$$

where

$$\begin{aligned} k_{rf} &= \frac{1}{r_{cutoff}^3}\left(\frac{\epsilon_{solvent}-1}{2\epsilon_{solvent}+1}\right) , \\ c_{rf} &= \frac{1}{r_{cutoff}}\left(\frac{3\epsilon_{solvent}}{2\epsilon_{solvent}+1}\right) . \end{aligned} \tag{2.13}$$

In the previous equations r_{cutoff} is the cutoff distance and $\epsilon_{solvent}$ is the dielectric constant of the solvent.

The "to dummy" and "from dummy" atom groups also require a special treatment for the intra-bonded molecular interactions. Indeed, the creation and destruction of atoms, leads to the creation and destruction of bonds, angles and dihedrals angles. Actually, the dihedral angle annihilations are not particularly difficult to implement. Indeed it is possible to set to zero the amplitude and phase of the angle dihedral to produce a null dihedral. On the other hand, set

to zero the force constants in the harmonic potentials related to the bond and angle terms in the mechanical potential (equation 2.8) could potentially produce unstable and noisy sampling. Indeed, for example, the progressively annihilation of the force constant in the potential bond term could lead to problematic bond shrinking which could produce difficult convergence problems. In order to overcome the problem it is interesting to observe that the free energy change between two thermodynamic states where the amplitude and the equilibrium constant of a harmonic potential are changed produces a variation equal to [92]:

$$\Delta F = \frac{k_B T}{2} \ln \frac{k_f}{k_i} \tag{2.14}$$

where, k_i and k_f are respectively the starting and the final force constant, k_B the Boltzmann constant and T the temperature. As a consequence, in an alchemical transformation where the creation or destruction of bond and angle terms are involved, free energy changes will annihilate if it is set $k_i = k_f$. This strategy was implemented in the developed code.

In the single topology method, a common topology between two mutant molecules is used to perform the alchemical transformation; the topology is morphed between the end states using the coupling parameter opportunely introduced into the system Hamiltonian. In this method, an important aspect is the determination of the topology parts that change between the end states. This is extremely important because it allows the detection of the different atom groups: "to dummy", "from dummy" and "hard". In order to detect structure changes between the end states, graph matching algorithms were used. These algorithms have been extensively applied in the analysis of chemical molecules and chemical reactions [93]. In these algorithms a molecule is treated as a mathematical graph. Briefly a graph G is made with a collection of N nodes (the atoms in a molecule), and a set E of arcs connecting pair of nodes $E = N \times N$ (the bonds in a molecule) [93]. Two nodes are adjacent if they are connected by one arc. A labeled graph is a graph where each node and arc has a label. A molecule can be effectively seen as a labeled graph where the atoms are labeled with atom names

and arcs with bond types. A subgraph of G is defined as a subset $P \subseteq N$ of nodes of G with a subset of its arcs $F \subseteq P \times P$ [93]. In addition, two subgraphs are isomorphic if there is a mapping between the nodes such as adjacent nodes in the first subgraph are mapped in adjacent nodes in the second subgraph [93]. A common subgraph of two graphs G_1 and G_2 consists of two subgraphs H_1 of G_1 and H_2 of G_2 such that H_1 is isomorphic to H_2 [93]. The Maximum Common Subgraph (MCS) of two graphs is the common subgraph which contains the largest possible number of arcs [93]. In order to detect changes between the end states of an alchemical transformation and detect the common topology, algorithms based on the detection of the MCS can be used. Most of the time these algorithms are bases on backtracking techniques [94] in conjunction with alpha-beta pruning [95]. These algorithms have a non polynomial complexity on the number of atoms and they are frequently applied between small molecules only such as ligands. Figure 2.8 shows the construction of the MCS between benzene and aniline molecules using an unlabelled MCS algorithm. In the implemented code, the MCS algorithm was not directly implemented but, an external computer program FeSetup [96] was used to generate an appropriate input file where all the mutations between the end states were detected, an therefore, to create the common topology.

The relative free energy code based on the FDTI and the single topology methods was developed extending Sire [74] through the use of the OpenMM APIs [75]. Sire is a molecular modelling framework developed in the C++ programming language. In order to simplify the programming to non-expert users, the Sire public class interface has been exposed to a more friendly programming language such as Python. Sire allows the creation of molecular systems using different approaches e.g. using Amber [97;98] topology files. It was originally developed to sample molecular systems by using the MMC algorithm and more recently by using MD (implemented by the same author of this thesis) through the use of the OpenMM APIs. These APIs are able to implement MD algorithms by using the modern parallel architectures present on the latest GPU and gain significant computational power compared to more conventional approaches. Furthermore, these APIs do not explicitly require the knowledge of dedicated GPU programming languages

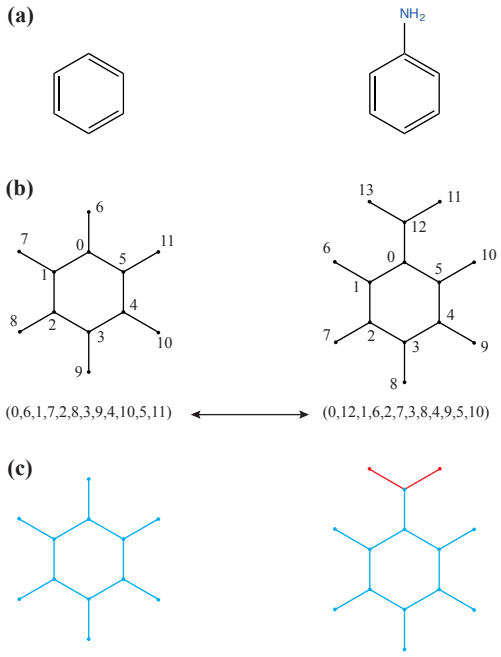


Figure 2.8: *(a,b) The benzene and aniline molecules are considered as two mathematical graphs. The nodes of each graph are labelled using integer numbers. The MCS algorithm finds the isomorphism between the largest common parts of the two chemical structures. In this case for example, the atom with index 6 in the benzene molecule is mapped into the atom with atom number 12 in the aniline molecule and so on. (c) The MCS graph is highlighted in light blue. The comparison of the MCS with the starting molecules allows the detection of the mutating parts in an alchemical mutation. In this case, these parts are highlighted in red.*

such as CUDA or OpenCL. Figure 2.9 illustrates the linking between Sire and its extension using OpenMM. In the relative free energy implementation OpenMM was used to code the mechanical potential described in the equation 2.8 with the use of soft core potential 2.10 and reaction field 2.12. The implementation of bonded and non-bonded terms was performed using custom potential expressions in OpenMM. In particular the inter-molecular potential applied to the different atom groups was implemented as follows:

- hard - hard interactions: Coulomb with reaction field and Lennard-Jones potential;

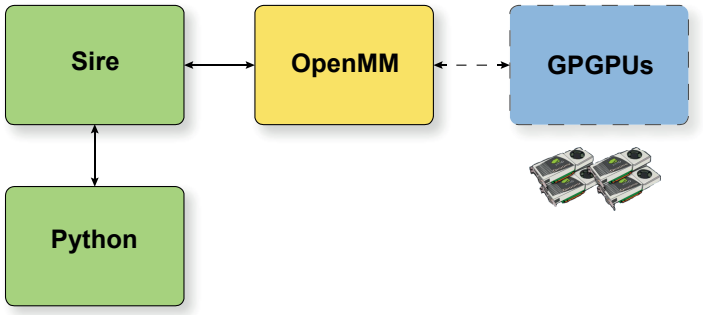


Figure 2.9: *The direct use of the Sire APIs could be difficult for scientists not expert in advanced programming because of the high abstracted interface implemented in C++. As a consequence, an easier front-end has been developed using Python wrappers. In Sire the system sampling is performed by using the MMC algorithm. Sire was linked with the OpenMM APIs to perform MD simulations using the capability of the latest GPUs. This has a dramatic impact on the speed-up of the MD simulations because of the huge number of cores and the high memory bandwidth present on these architectures.*

- to dummy - to dummy interactions: soft core potential with reaction field;
- from dummy - from dummy interactions: soft core potential with reaction field
- hard - to dummy: soft core potential with reaction field;
- hard - from dummy: soft core potential with reaction field;
- to dummy - from dummy: soft core potential with reaction field.

Figure 2.10 reports a summary of the inter-molecular potential used. The 1-4 intra-molecular non bonded interactions were implemented with custom bonded OpenMM expressions using the same schema represented in Figure 2.10. This potential was not applied to atom particles separated by one or two covalent bonds (1-2, 1-3 interactions). The intra-molecular bonded terms were implemented by using custom OpenMM potential expressions for the bond, angle and dihedral terms.

In order to test and calibrate the implementation, single point energy calculations were performed on different systems for selected values of the coupling

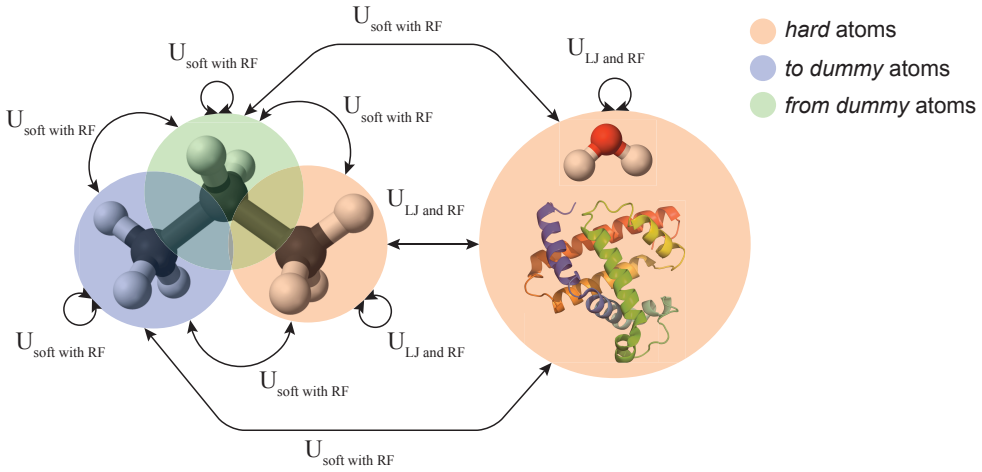


Figure 2.10: *The implemented inter-molecular interactions between a mutant solute and a system composed by solvent molecules and a protein. The inter-molecular interactions between "hard" atoms are evaluated using the standard LJ potential (equation 1.17) and reaction field potential (equation 2.11) and named in figure as* $U_{LJ\ and\ RF}$*. The other inter-molecular interactions are evaluated by using soft core potential with reaction field (equation 2.10 where the Coulomb term is replaced by equation 2.12) and named in figure as* $U_{soft\ with\ RF}$ *.*

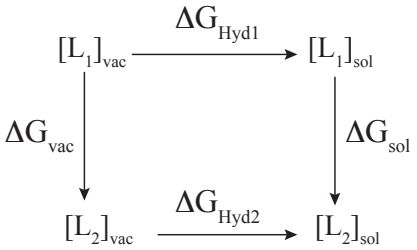
parameter. The total potential energy for different alchemical mutations in vacuum was computed by using the implemented code in OpenMM and then compared with the same potential energy implemented in Sire only without using OpenMM. The OpenCL platform was selected in OpenMM on GPU with mix precision while the calculation in Sire were performed on CPU in double precession. A total of 100 mutations involving different polar and non polar molecule groups were considered and the software package FEsetup [96] was used to set the different systems and to generate the common topology used to perform the alchemical mutation and table (2.1) reports all the selected mutations. The point atomic charges were assigned by using the Amber module Antechamber [97;98], selecting the AM1-BCC method [28] and GAFF [97;98] was used for the generation of the other force field parameters. The total system potential energy was computed for eleven values of the coupling parameter λ equally spanned over the range [0,1]: (0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0). The statistical agreement between the different window values was assessed evaluating the absolute error between the calculated potential energy in Sire and OpenMM. On average the Mean Unsigned Error (MUE) recorded considering all the mutations was $\simeq 10^{-5}$, which is in agreement with similar tests used to compare CPU vs GPU potential energy errors [99]

Table 2.1: *The total potential energy for eleven windows in 100 alchemical mutations was calculated with OpenMM and compared to the same potential implemented in Sire only. The statistical agreement for each system was assessed evaluating the absolute error for each single point energy calculation between Sire and OpenMM. On average the overall MUE* $\simeq 10^{-5}$

System number	Transformation	System number	Transformation
1	acetamide to acetone	51	methylbenzene to benzene
2	acetamide to methylethene	52	methylbenzene to methane
3	acetone to acetamide	53	methylbenzene to phenol
4	acetone to dimethylether	54	methylethene to acetamide
5	acetone to formaldehyde	55	methylethene to formaldehyde
6	acetylbenzene to aniline	56	methylethene to methylacetylene
7	acetylbenzene to methylacetatebenzene	57	methylfuran to methane
8	aniline to acetylbenzene	58	methylfuran to methylpyrrole
9	aniline to methylbenzene	59	methylfuran to methylthiene
10	aniline to nitrobenzene	60	methylfuran to methyltriazole
11	aniline to pyridine	61	methylfuryltriazole to methyltriazole
12	benzene to chlorobenzene	62	methylindole to methane
13	benzene to cyclopropylbenzene	63	methyloxadiazole to methyloxazole
14	benzene to ethynebenzene	64	methyloxadiazole to methyltriazole
15	benzene to methylbenzene	65	methyloxazole to methyloxadiazole
16	benzene to nitrilebenzene	66	methyloxazole to methylpyrrole
17	benzene to pyridine	67	methyloxazole to methyltriazole
18	benzene to triazine	68	methylphenyltriazole to methyltriazole
19	chlorobenzene to benzene	69	methylpyrrole to methylfuran
20	chlorobenzene to nitrilebenzene	70	methylpyrrole to methyloxazole
21	chlorobenzene to phenol	71	methylpyrrole to methylthiene
22	cyclopropylbenzene to benzene	72	methylthiene to methylfuran
23	dimethylether to acetone	73	methylthiene to methylpyrrole
24	dimethylether to dimethylsulfide	74	methyltriazole to methylfuran
25	dimethylether to methanol	75	methyltriazole to methylfuryltriazole
26	dimethylether to propane	76	methyltriazole to methyloxadiazole
27	dimethylsulfide to dimethylether	77	methyltriazole to methyloxazole
28	dimethylsulfide to propane	78	methyltriazole to methylphenyltriazole
29	ethane to methane	79	neopentane to ibutane
30	ethane to methanol	80	neopentane to methane
31	ethane to propane	81	nitrilebenzene to benzene
32	ethynebenzene to benzene	82	nitrilebenzene to chlorobenzene
33	ethynebenzene to nitrilebenzene	83	nitrilebenzene to ethynebenzene
34	formaldehyde to acetone	84	nitrobenzene to aniline
35	formaldehyde to methylethene	85	nitrobenzene to pyridine
36	ibutane to neopentane	86	nitrobenzene to triazine
37	ibutane to propane	87	phenol to chlorobenzene
38	methane to ethane	88	phenol to methylbenzene
39	methane to methanol	89	propane to dimethylether
40	methane to methylbenzene	90	propane to dimethylsulfide
41	methane to methylfuran	91	propane to ethane
42	methane to methylindole	92	propane to ibutane
43	methane to neopentane	93	propane to methane
44	methane to propane	94	pyridine to aniline
45	methanol to dimethylether	95	pyridine to benzene
46	methanol to ethane	96	pyridine to nitrobenzene
47	methanol to methane	97	pyridine to triazine
48	methylacetatebenzene to acetylbenzene	98	triazine to benzene
49	methylacetylene to methylethene	99	triazine to nitrobenzene
50	methylbenzene to aniline	100	triazine to pyridine

2.4 Relative Hydration Free Energy calculation. Ethane to Methanol a case study

The relative free energy implementation was tested on many systems. In this section, results related to the calculation of the relative hydration free energy between the ethane and methanol molecules are reported. The alchemical path used to perform the calculation is shown in Figure 2.11. In this case, it is necessary



$$\Delta\Delta G_{Hyd} = \Delta G_{Hyd2} - \Delta G_{Hyd1} = \Delta G_{sol} - \Delta G_{vac}$$

Figure 2.11: *The relative free energy of hydration between two molecules L_1 and L_2 can be computed performing two alchemical mutations. In the first simulation the molecule L_1 is mutated into the molecule L_2 (ΔG_{vac}) and in a second simulation the same mutation takes place in a solvent environment (ΔG_{sol}). The differences between these two free energy changes equates the relative hydration free energy: $\Delta\Delta G_{Hyd} = \Delta G_{sol} - \Delta G_{vac}$.*

to perform two alchemical mutations between the two molecules respectively in an aqueous environment and in vacuum. The software package FEsetup[96] was used to set the systems and generate the common topology used to perform the alchemical mutations. The point atomic charges were assigned by using the Amber module Antechamber[97;98] selecting the AM1-BCC method[28]. GAFF[97;98] was used for the generation of the other force field parameters and the system was also solvated in a box of water selecting TIP3P as water model by using the Amber module LEaP[97;98]. Before the production run, the system was minimized for 100 cycles by using the steepest descent method and equilibrated a 298K and 1 atm pressure for 200 ps - 2 fs time step by using the Amber module Sander[97;98]. In this process harmonic restraint were set on the common topology

by using a restraint force constant of 10 kcal mol^{-1} Å^{-2} and selecting the starting atom positions as restraint positions. The hydrogen bonds were constrained to their equilibrium distances in this stage. In order test the implementation the system sampling was performed using MD through OpenMM and using the standard MMC in Sire. Sixteen values of the coupling parameter λ were selected in the range [0,1]: (0.0000, 0.0109, 0.0432, 0.0955, 0.1654, 0.2500, 0.3455, 0.4477, 0.5523, 0.6545, 0.7500, 0.8346, 0.9045, 0.9568, 0.9891, 1.0000). These values were calculated by using the Chebyshev nodes to improve the numerical stability of regression techniques used to estimate the TI integral. Indeed recent studies have shown that the use of polynomial regression using equidistant values could lead to convergence problems when the interpolation is performed with high polynomial orders[100]. Chebyshev nodes have been proven to mitigate this issue especially in the TI method when the interpolated curve is not smooth enough. The coupling parameters λ were generated using the following equation derived from the Chebyshev nodes[100]:

$$\lambda = \frac{1}{2}|\cos(\psi\pi) - 1| \tag{2.15}$$

where $\psi \in [0, 1]$. The MD production run was performed for 5 ns constraining all the bonds and selecting a time step of 2 fs. During the simulation 25×10^4 gradient values were collected and used to compute the gradient for each window in the solvent and in the vacuum simulation. In addition in the solvent simulations the pressure and the temperature were kept constant respectively using the Monte Carlo Barostat algorithm[101;102], setting the Monte Carlo frequency to 25 steps, and using the Andersen thermostat[82] with a collision coefficient of 1/ps. The coulomb power and the delta shift soft-core parameters were respectively set to 0 and 2. All the bonds containing hydrogens were constrained to their equilibrium distance and the non-bonded interactions were evaluated by using an atom based cut off scheme setting the cutoff distance to 10 Å . In the solvent simulation, the electrostatic interactions were calculated by using reaction field with the medium dielectric constant set to the water dielectric constant ($\epsilon_{solvent} = 78.3$). The sampling by using the MMC method was performed in the solvent state selecting $200 \times 5 \cdot 10^5$ moves. In each move the solvent molecules were allowed to

perform a maximum translation of 0.15 Å and a maximum rotation of 15 degrees. The solute was allowed to move with a maximum translation of 0.192 Å and a maximum rotation of 15 degrees. In addition the maximum volume move was set to 360 $Å^3$. For the vacuum state by using MMC method the number of internal Monte Carlo moves was set to 200×1000. The simulation in vacuum and solvent were repeated three times for both sampling methods and the uncertainties were calculated as standard deviation of the mean over the three independent runs. Figure 2.12 reports the free energy gradient variations along the alchemical mutations in vacuum and solvent sampling the system by using MD and MMC. The free energy gradient values were integrated using a seven order polynomial

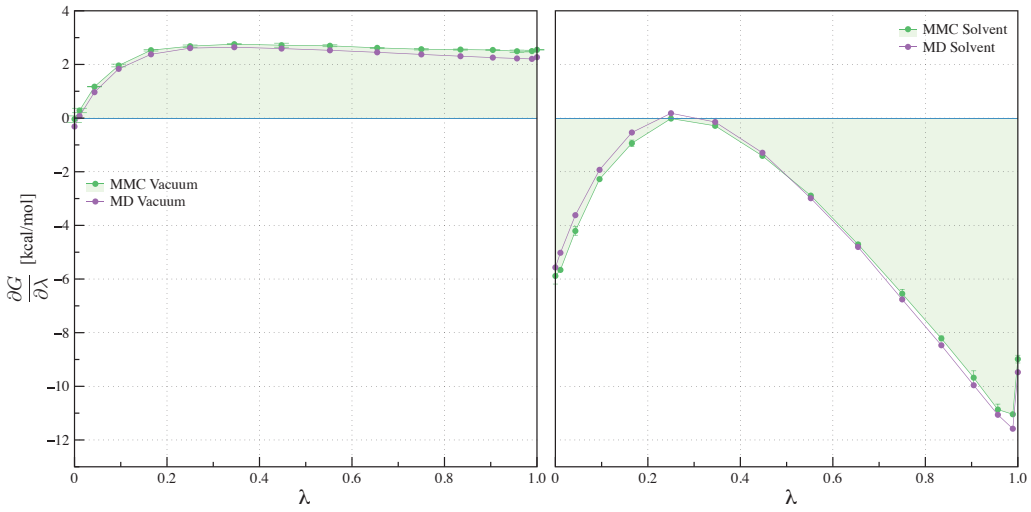


Figure 2.12: *Free energy gradient evaluated for the vacuum and solvent mutations using MD and MMC. The free energy gradient was computed for 16 windows selected in the range [0,1]. The difference between the highlighted areas represents the relative hydration free energy for the MMC case only.*

regression and the computed values for the relative hydration free energies by using MD and MMC were respectively -6.270 ± 0.003 kcal/mol and -6.49 ± 0.03 kcal/mol. The experimental value calculated by using the solvation free energy database reported by Mobley [103] was: -6.93 kcal/mol.

The relative ethane to methanol hydration free energy calculation was historically simulated for the first time by using the single topology method, FEP and

MMC by Jorgensen et al.[53] and nowadays it is often used as horse test for new single topology implementations. It is interesting to observe that the calculated value was -6.93 kcal/mol in the original paper, which is in very close agreement with the experimental value just using few windows. The discrepancy with the calculated values here could be blamed to the force field. In the Jorgensen paper the calculation was performed by using the OPLS force field to parameterise the ethane and methanol molecules while TIP4P water model was selected for the solvent and it could be the luckily reason for the experimental discrepancy.

Our results show that the hydration free energies computed by using MD and MMC are not consistent in error bars. Theoretically, in the limit of infinite sampling time the results have to be the same but, practically, this often does not happen. It is well known, that one of the main differences between the methods is related to the efficiency in exploring the configurational space. In the MMC method, the system is evolved using random moves and therefore the method is ergodic by construction. On the other hand, by using MD to generate moves the system tends to move in regions of configuration space with lower energy and, therefore climbing high energy barriers could be very problematic. In the considered system, due to the relatively small number of particles and the selected setting to perform the simulations the sampling should not be the cause of the discrepancy. It is possible that this is related to the thermostat and/or barostat schemes used and their setting to perform the MD simulations, which did not correctly sample the NPT ensemble.

2.5 Absolute Hydration Free Energy calculation. 1,2-Dichloroethane a case study

A flavour of dual topology method was also developed to calculate absolute hydration free energy using the alchemical path illustrated in Figure 2.3 (a). The implementation was applied to the calculation of the absolute hydration free energy of 1,2-dichloroethane molecule. In the selected alchemical path, the intermolecular interactions between the solute and solvent were progressively switched on by using the soft core potential. The switching was performed in two different

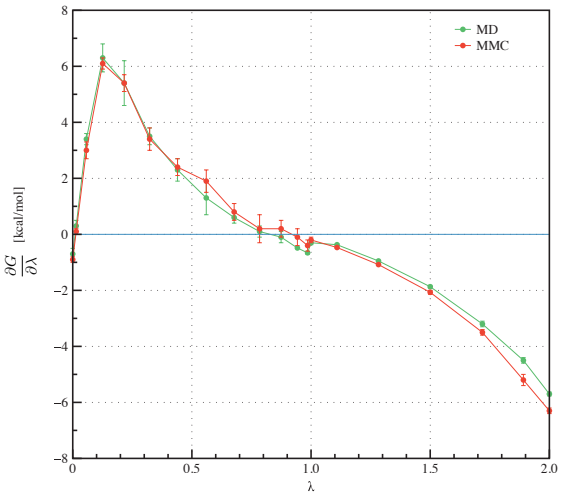
stages. In the first stage the VdW interactions were turned on and subsequently, in a second stage, the Coulomb interactions were turned on. In particular, twenty coupling parameters λ were selected in the range $[0, 2]$. In this range when the λ values were varying between $[0, 1]$ the VdW forces related to the soft-core term, were progressively switched on while the Coulomb forces were kept off (first stage). For higher values of $\lambda \in (1, 2]$ the VdW terms were kept on and coulomb forces were progressively switched on (second stage). This approach was used to avoid steric clashes between the solute and solvent atoms with opposite charges at the beginning of the simulation. The 1,2-dichloroethane molecule was parameterised using GAFF[97;98] and the point atomic charges were assigned using the Amber module Antechamber[97;98] selecting the AM1-BCC method[28]. In addition, TIP3P water model was used for the solvent. Before the production run the system was minimized for 100 cycles by using the steepest descent method and equilibrated at 298K and 1 atm pressure for 200 ps - 2 fs time step by using the Amber module Sander[97;98]. In this process harmonic restraint were also set on the solute molecule by using a restraint force constant of 10 kcal mol^{-1} $Å^{-2}$ and selecting the starting solute atom positions as restraint positions. The system was sampled by using two different methodologies: MMC and MD. For each method the simulations were repeated three times and statistical uncertainties were calculated using the standard deviation of the mean. The MD simulations were performed for a total time of 1 ns using a time step of 1 fs. During the production run using MD all bonds and angles containing hydrogen atoms were constrained and the temperature was set to 298 K while the pressure was kept constant at 1 atm. The pressure and the temperature were controlled respectively using the Monte Carlo Barostat algorithm[101;102], setting the Monte Carlo frequency to 25 steps, and using the Andersen thermostat[82] with a collision coefficient 1/ps. The coulomb power and the delta shift soft-core parameters were respectively set to 0 and 2. All the bonds containing hydrogens were constrained to their equilibrium distance and the non-bonded interactions were evaluated by using an atom based cut off scheme setting the cutoff distance to 10 Å . The electrostatic interactions were calculated by using reaction field with the medium dielectric

constant set to the water dielectric constant ($\epsilon_{solvent} = 78.3$). During the MD sampling $50 \cdot 10^4$ gradient values were collected and used to calculate the ensemble average. The sampling by using the MMC method was performed selecting 100×10^6 moves. In each move the solvent molecules were allowed to perform a maximum translation of 0.15 Å and a maximum rotation of 15 degrees. The solute was only allowed to have a maximum rotation of 15 degrees. In addition the maximum volume move was set to 122.25 Å^3. Twenty windows were selected in the range [0,2]: (0.0000, 0.0145, 0.0573, 0.1257, 0.2160, 0.3227, 0.4397, 0.5603, 0.6773, 0.7840, 0.8743, 0.9427, 0.9855, 1.0000, 1.1090, 1.2830, 1.5000, 1.7170, 1.8910, 2.0000). These values were calculating by using the Chebyshev technique previously described Chebyshev equation (2.15). With the aim of further improving the integral calculation, 14 out of 20 points were selected in the λ range $[0, 1]$. In this range the average gradient undergoes higher variations because of the relative short distances between water molecules and the 1,2-dichloroethane molecule with significant contributions related to the VdW forces. The TI integral was calculated interpolating the gradient data with a polynomial regression of order seven. In addition, trapezium integration was also carried out and compared with the polynomial integration. The results related to the evaluation of the gradient for both methods are shown in Figure 2.13. Table 2.2 reports a comparison of the calculated hydration free energy sampling the system by using the MD and MMC methods and Figure 2.14 shows a comparison between the convergence time

Table 2.2: *Values of free energy of hydration related to the 1,2-dichloroethane molecule at 298K sampling the system by using the MD and MMC methods. The numerical integration was carried out using a polynomial regression of order seven and using the trapezium integration quadrature rule. Data is reported in kcal/mol along with the experimental data. Three independent runs were performed for MD and MMC and uncertainties were calculated propagating the standard deviation of the mean of the recorded binding affinities.*

	MD	**MMC**
Polynomial Regression	-0.18 ± 0.06	-0.3 ± 0.1
Trapezoid	-0.15 ± 0.07	-0.3 ± 0.1
Experimental data	-0.17 ± 0.01	

of the cumulative gradient by using both methods. The agreement between the



	MD	MMC
λ	$\frac{\partial G}{\partial \lambda} \pm \delta$	$\frac{\partial G}{\partial \lambda} \pm \delta$
0.0000	-0.7 ± 0.2	-0.9 ± 0.1
0.0145	0.3 ± 0.2	0.1 ± 0.1
0.0573	3.4 ± 0.2	3 ± 0.3
0.1257	6.3 ± 0.5	6.1 ± 0.2
0.2160	5.4 ± 0.8	5.4 ± 0.3
0.3227	3.5 ± 0.3	3.4 ± 0.4
0.4397	2.3 ± 0.4	2.4 ± 0.3
0.5603	1.3 ± 0.6	1.9 ± 0.4
0.6773	0.6 ± 0.2	0.8 ± 0.3
0.7840	0.1 ± 0.2	0.2 ± 0.5
0.8743	-0.1 ± 0.2	0.2 ± 0.3
0.9427	-0.486 ± 0.007	-0.1 ± 0.3
0.9855	-0.66 ± 0.05	-0.4 ± 0.2
1.0000	-0.30 ± 0.07	-0.2 ± 0.1
1.1090	-0.37 ± 0.03	-0.47 ± 0.02
1.2830	-0.95 ± 0.02	-1.08 ± 0.02
1.5000	-1.87 ± 0.05	-2.07 ± 0.07
1.7190	-3.2 ± 0.1	-3.5 ± 0.1
1.8910	-4.5 ± 0.1	-5.2 ± 0.2
2.0000	-5.70 ± 0.07	-6.3 ± 0.1

Figure 2.13: *Values of the free energy gradient per each coupling parameter λ used for the MD and MMC methods. The values are expressed in kcal/mol. The gradient values are the arithmetic average over three runs and δ represents the related standard deviation of the mean.*

two methodology was quite satisfactory and assessed calculating the determination coefficient: $R^2 = 0.99$. It is interesting to observe that the binding affinity calculated with MD method was closer to the experimental value compared to MMC. Figure 2.13 shows that the difference is mainly originated when the 1,2-dichloroethene molecule is nearly fully decoupled from the solvent environment. As previously described, it is notorious that the dual topology method is hard to converge especially at the end states of alchemical mutations and longer sampling is usually required. Figure 2.14 shows convergence problem for one of the selected window for the MMC method compared to the MD method and this could explain the discrepancy between the two approaches in this particular calculation.

2.6 Chapter Conclusions

In this chapter was presented an implementation of the single topology alchemical method and the FDTI method used in conjunction to calculate relative binding affinities. The implementation was performed merging two pieces of software Sire and the OpenMM APIs. The former is a molecular modelling framework

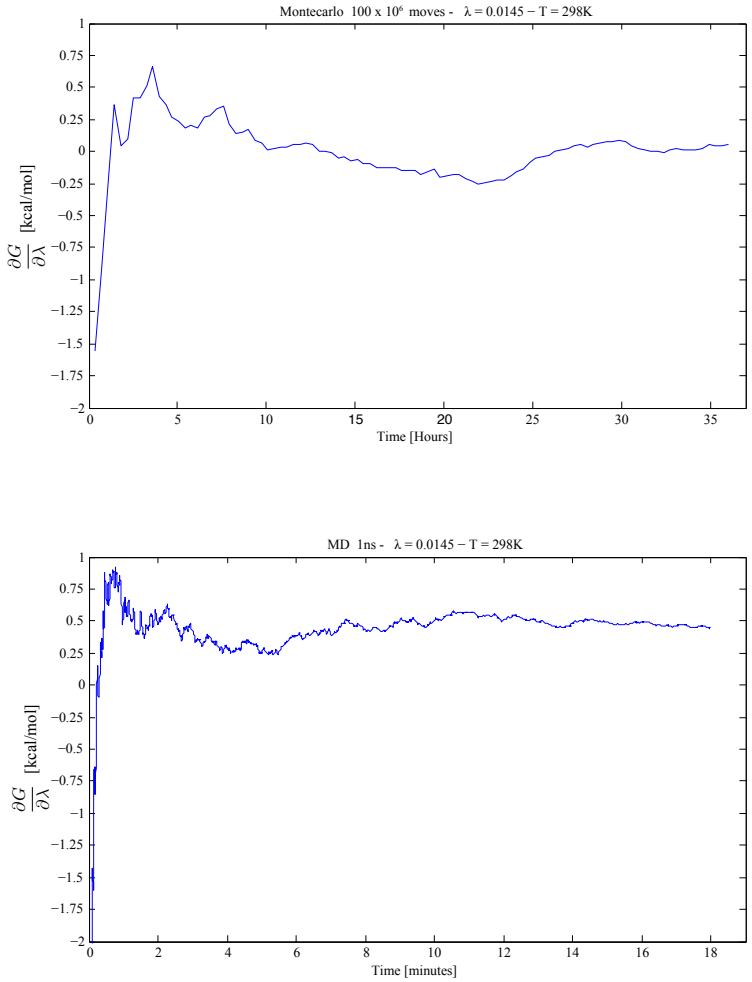
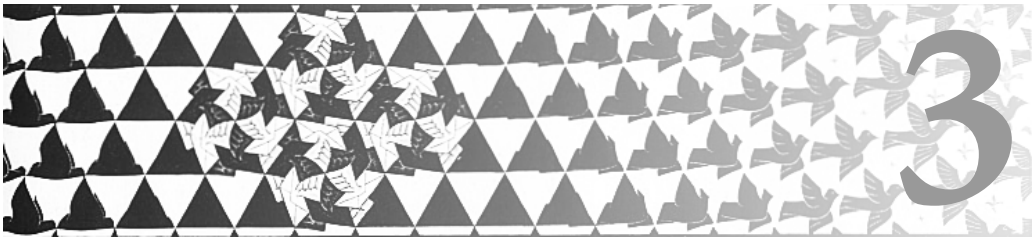


Figure 2.14: *Variations of the cumulative gradient at the temperature of 298K for the value of the coupling parameter* $\lambda = 0.0145$ *versus the total simulation time. The sampling of* 100×10^6 *MMC moves required about 35 hours for the case study. On the other hand, the MD time required to perform 1 ns simulation was 18 minutes. The final convergent value is different between the two methods and overall the MMC sampling seems to be more problematic than the MD sampling. This could suggest that longer sampling is required for the MMC method for the selected window.*

extremely flexible in bio-molecular modelling, definition and editing of molecular parameters and implementation of new molecular simulation methods. On the other hand, the OpenMM APIs allows the simulation of molecular system by using MD accelerated by using the latest piece of hardware such as GPU. In order to test the code, the implemented potential energy function by using the OpenMM APIs was compared to the same potential energy function implemented in Sire only and single point calculation energy were performed on 100 alchemical mutations. Results were in excellent agreement with a MUE error of $\simeq 10^{-5}$. Subsequently, the relative hydration free energy of ethane to methanol was calculated by using the implemented code and MMC. The results were in good agreement with the experimental value ($<$ 1 kcal/mol) although, the MMC result was closer to the experimental value compared to the MD based implementation. The discrepancy could be caused by a not correct sampling of the selected NPT ensemble. Finally, the absolute hydration free energy of 1,2-dichloroethene was calculated. In this case, the MD based implementation result was in excellent agreement with the experimental data compared to the MMC method. The problem seems to be related to the sampling of few windows closer to the fully decoupled end state that are not sampled enough by using the MMC method. In Chapter four the code will be validate on a more significant bio-molecular system such as the Thrombin protein.

"There are two possible outcomes: if the result confirms the hypothesis, then you've made a measurement. If the result is contrary to the hypothesis, then you've made a discovery"

— Enrico Fermi



Influence of molecular flexibility on conformational equilibrium

3.1 Introduction

For a long time chemists and biochemists considered molecules just in terms of their atomic constituents and two dimensional structure. The importance of three dimensional conformations was only discovered later on by semiochemists, when they became interested in the interactions between fragrant substances with receptors [104]. This led to the notion of a molecular shape that reflects a given arrangement of the atoms of a molecule in space. In general, different molecular conformations need to be considered to explain many chemical-physical properties that cannot be explained with just a single conformation. Conformers are stereoisomers in dynamic equilibrium that could differ from each other through stretching and bending of bonds or rotations around covalent

bonds[105]. Classes of molecules that may readily adopt different conformers are often named "flexible" molecules. Molecular flexibility has been studied for long time because of its importance in many contexts such as polymers, protein folding or protein-ligand binding. For instance, in the latter, ligand functional groups frequently need to adopt specific spatial arrangements to match the complementary shape of the protein binding site.

In the next paragraphs of this chapter, an experimental and computational study performed on the influence of molecular flexibility on the preferred conformations of a set of related molecules will be detailed. The focus was to collaborate with the experimental group of Dr Cockroft (University of Edinburgh) to reproduce *in-silico* the experimental data derived from NMR measurements. As the computational models used were relatively inexpensive, the influence of force field parameters on the resulting conformational equilibria was investigated. Such studies are important to validate the accuracy of forcefields used in simulations of large biological molecules.

3.2 The experimental systems

With the aim of studying the effect of molecular flexibility on conformational equilibria, a set of molecules with two main moieties linked together by flexible chains were experimentally synthesised. The two groups named here α and β were designed to interact with each other by forming intra-molecular interactions in solution. These molecules can adopt different conformers as schematically illustrated in Figure 3.1. Many molecules were synthesised with different chain lengths and Figure 3.2 represents the whole set of synthesised molecules. In each molecule the α group is a 4-fluorophenol group, β an amide group and the linker chain an alkyl chain. The hydrogen of OH in the 4-fluorophenol group is able to form an intra-molecular hydrogen bond with the oxygen atom in the amide group. In order to investigate how this intra-molecular hydrogen bond was affected by the chain length, each linker molecule was solvated in a solution made of chloroform and, a tributylphosphine at selected concentrations. Following the experimental setup, the thermodynamic process can be represented as illustrated

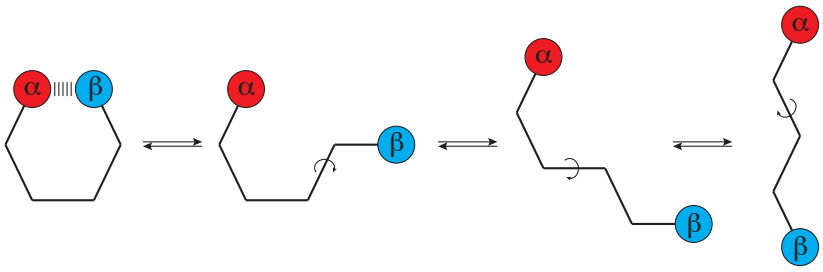


Figure 3.1: *Schematic representation of the synthesised molecules. In solution each molecule can adopt different conformations and can also form an intra-molecular hydrogen bond between the groups α and β. The chain length has a significant impact on the stability of the intra-molecular hydrogen bond conformer.*

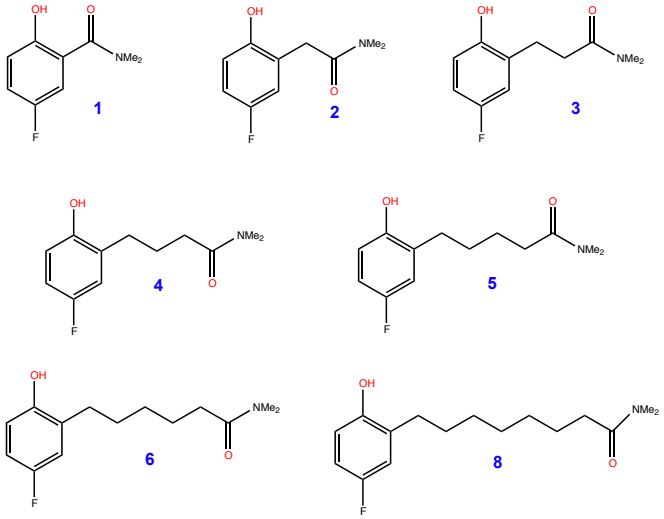


Figure 3.2: *The set of synthesised molecules. The alkyl amide substituent contains 1, 2, 3, 4, 5, 6 and 8 carbon atoms excluding the two terminal methyl group. The chain links a 4-fluorophenol group with an amide group. The two groups are able to form an intra-molecular hydrogen bond between the hydrogen of OH in the 4-fluorophenol group and the oxygen of the amide carbonyl group.*

in Figure 3.3 in solution. The whole process presents three different thermody-

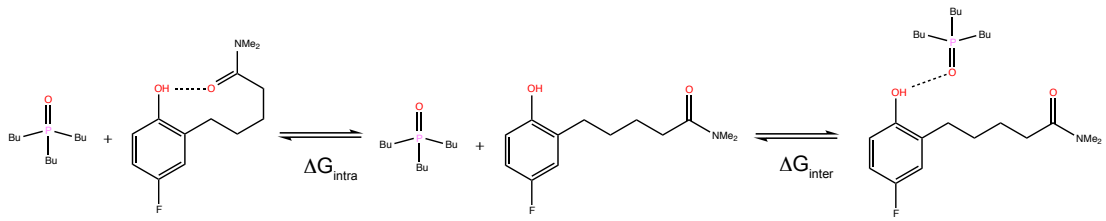


Figure 3.3: *The thermodynamic process in solution can be represented by three thermodynamic states. Each linker molecule is able to form intra- and inter-molecular hydrogen bonds. The intra-molecular hydrogen bond defines the folded and unfolded states. The inter-molecular hydrogen bond is formed with the tributylphosphine oxide molecule.* ΔG_{inter} *and* ΔG_{intra} *are respectively the inter- and intra- molecular complexation free energies.*

namic states in dynamic equilibrium. A first state named the "unfolded" state where the linker molecule is not forming either intra- or inter- molecular hydrogen bond with itself or another solute. A second state named the "folded" state where the linker molecule can form intra-molecular hydrogen bond interactions. Finally, a third thermodynamic state where the linker molecule can form an hydrogen bond between the hydrogen of OH in the 4-fluorophenol group and the oxygen of tributylphosphine oxide molecule. Experimentally, the change in chemical shift of the ^{31}P signal of the tributylphosphine oxide was measured by using NMR titration and, just one chemical shift was observed. Therefore, the process illustrated in Figure 3.3 seems to be a fast equilibrium process. The observed strength of the tributylphosphine oxide bound complex is in direct competition with the intermolecular complex and, therefore, the measured energy of complexation ΔG_{obs} is:

$$\Delta G_{obs} = \Delta G_{inter} - \Delta G_{intra} \ . \tag{3.1}$$

In the previous equation, ΔG_{intra} is the intra-molecular complexation energy required to form the intra-molecular hydrogen bond and ΔG_{inter} is the inter-molecular complexation energy required to form the inter-molecular hydrogen bond. In order to indirectly measure ΔG_{intra}, the direct measurement of the

ΔG_{inter} contribution was performed by using NMR titration. To do this a molecule similar to those depicted in Figure 3.2 was synthesised but lucking the amide group. In this case, the thermodynamic process in solution can be described as illustrated in Figure 3.4. The inter-molecular complexation energy ΔG_{inter} is

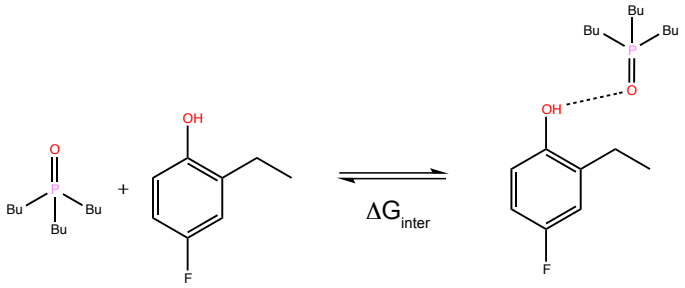


Figure 3.4: *The amide group present in the original molecules was removed. In this case, the thermodynamic process can be represented by two thermodynamic states. This new molecule can only form inter-molecular hydrogen bond with the tributylphosphine oxide molecule. The inter-molecular complexation energy ΔG_{inter} was measured by using NMR titration.*

independent from the alkyl chain length and experimentally its measured value was: $\Delta G_{inter} = -2.15$ kcal/mol. Figure 3.5 reports the overall measured complexation energy ΔG_{obs} and the indirect estimation of ΔG_{intra} for the different linker molecules. The intra-molecular complexation energy can be decomposed into an entropic and an enthalpic component. The former is associated to the entropy related to the different conformations produced by rotations around the alkyl chain and the latter is the change due to the enthalpy related to the intra-hydrogen bond formation. It is clear from the experimental results that increasing the alkyl chain length produces an increase in ΔG_{intra} and, therefore, an increase in the entropic component.

3.3 The computational systems

The main goal of this research project was to calculate in-silico the intra-molecular complexation energy ΔG_{intra} for the different linker molecules. In particular, this study highlighted the importance of the force field parameterisations. The se-

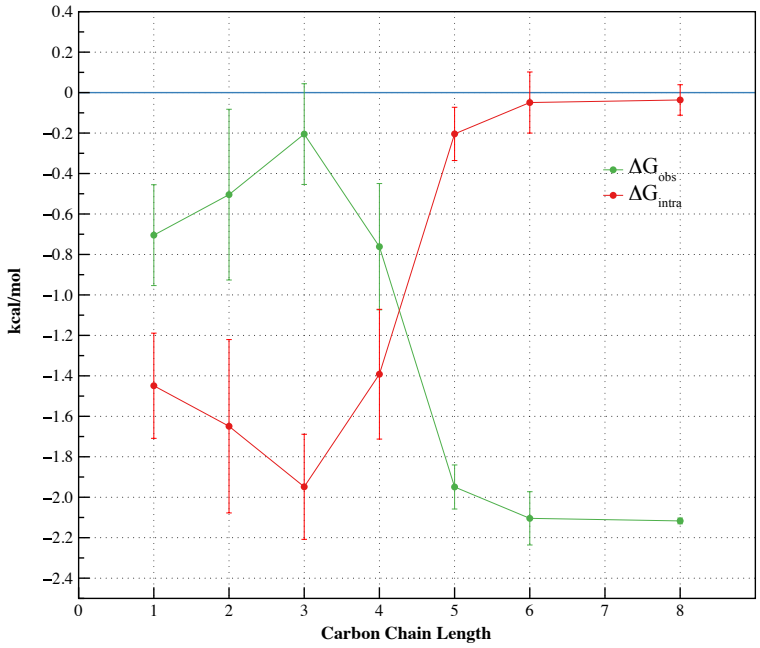


Figure 3.5: *The overall experimental complexation energy ΔG_{obs} and the related intra-complexation energy ΔG_{intra}.*

lected molecular systems are indeed relatively small to enable extensive sampling and precise calculations of ensemble averages at low compuatational cost. This permits the evaluation of different parameter sets in a timely manner. The computation of the intra-molecular complexation energy can be performed by using the thermodynamic process illustrated in Figure 3.6. In Chapter one, the FEP and

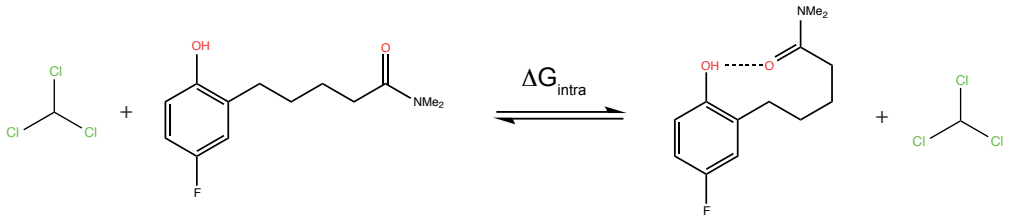


Figure 3.6: *The thermodynamic process used to calculate ΔG_{intra}.*

TI methods were introduced to calculate free energy changes. However, the histogramming technique is another popular method used to calculate a free energy

change ΔG. This approach relies of the following thermodynamic equation:

$$\Delta G = -k_b T \ln \frac{P_1}{P_0} \ , \tag{3.2}$$

where k_b is the Boltzmann constant, T the temperature and P_1 and P_0 are respectively the probabilities to find the system in the final and starting thermodynamic state. In this particular case, P_1 and P_0 are respectively the probability of the "folded" and "unfolded" state. From a computational point of view, it is possible to estimate these probabilities by sampling the systems conformations by molecular simulations and, by computing the relative frequencies of the two states recorded along the simulation. In the limit of sufficiently large sample size, the relative frequencies approach the equilibrium state probabilities. However, this approach requires to discriminate between folded and unfolded state by using an appropriate criteria. In the examined systems the folded state is related to the intra-molecular hydrogen bond formation. A reasonable threshold distance to observe hydrogen bond is 2.5 Å between acceptor (oxygen atom in amide group) and the hydrogen atom of the donor (phenolic hydroxyl) and it will be assumed from now on. Although the systems are not computationally demanding in this case, the force field parameterization was quite crucial to reproduce the experimental results. Indeed, the experimental set up solvated the different linker molecules in a solution made of chloroform and tributylphosphine oxide molecule. The evaluation of the intra-complexation energy using the thermodynamic process in Figure 3.6 requires the modeling of the system in a solution of pure chloroform. In the molecular simulation literature, this solvent is quite unusual and in principle, it could have a significant impact on the linker molecule parameterization. In particular the atomic charge calculations. In order to find a suitable protocol to perform MD simulations on each linker molecule, two atomic charge computation methods were considered: AM1-BCC[28] and the Charge Model 5 (CM5)[106]. The AM1-BCC method is used to quickly produce high quality atomic charges for computer simulation of organic molecule in polar media[28]. The method emulates the quantum electrostatic potential derived using the HF/6-31G* level of

theory. This theory level has been shown to reproduce hydration free energies of organic molecules with good accuracy and, therefore, it is adequate for aqueous simulations[28]. In the AM1-BCC method the atomic charge $q_j^{AM1-BCC}$ of an atom j is evaluated as follows[28]:

$$q_j^{AM1-BCC} = q_j^{AM1} + q_j^{BCC} \ , \tag{3.3}$$

where q_j^{AM1} is the charge evaluated by using the AM1[107] semi-empirical method and q_j^{BCC} is the correction charge calculated by using the Bond Charge Correction method BCC[28]. In this method the charges are evaluated by using the equation[28]:

$$q_j^{BCC} = \sum_{\alpha}^{\gamma} T_{j\alpha} p_\alpha \ , \tag{3.4}$$

where γ is the total number of bond types present in the molecule, $T_{j\alpha}$ is the bond connectivity template matrix and p_α is the bond charge correction term. In the AM1-BCC method, the p_α parameters are estimated fitting a training set of more than 2700 molecules with the electrostatic potential generated by using HF/6-31G* level theory.

CM5 is a computational method used to calculate atomic charges designed to accurately reproduce the dipole moment of a molecule by mapping the Hirshfeld[108] population charges onto a new set of charges[106]. The model can be applied to charged and uncharged molecule in gas or in solution. CM5 is a class IV charge model i.e. the atomic charge are calculated to accurately reproduce charge-dependent observable e.g. the dipole moment by using experimental data or high level quantum mechanics calculations. CM5 is the successor of the CMx models that were introduced to overcome the classes I,II and III charge model problems such as the charge dependence on the molecular system orientations, the dependence on the theory level and the selected basis set[106]. CM5 is based on the Hirshfeld population, which is less sensitive to basis set size and the basis set choice compared to the CMx models[106]. The CM5 charge q_k^{CM5} is calculated

by using the following equation[106]:

$$q_k^{CM5} = q_k^{HPA} + \sum_{k' \neq k} T_{kk'} B_{kk'} \ , \tag{3.5}$$

where q_k^{HPA} is the Hirshfeld charge, $B_{kk'}$ are constant values function of tabulated atomic covalent radius and $T_{kk'}$ are the parameters to optimise. In the optimisation process the MG3S basis set is used along with a training set of 614 molecular configurations with ether experimental reference dipole moments (388 data) or, theoretical reference dipole moments (226 data)[106]. The latter are density based dipole moments averaged over five theoretical methods[106]. The charges produced by the CM5 model are used for force field parameterisation in MD simulations, for computing solvation free energy and for generating realistic potentials and multipole moments[106].

In order to estimate the impact of the selected charge methods on the calculations the intra-molecular complexation energy was initially calculated for the linker molecule with three carbon bonds along the alkyl carbon chain. The linker molecule was modelled by using Schrödinger Maestro[109] and the following protocols named here AM1-BCC and CM5 were set to calculate the atomic charges and to assemble the relative molecular systems:

- AM1-BCC protocol
  - The atomic charges were assigned by using the Amber module Antechamber[98] and selecting the AM1-BCC method;
  - GAFF[97;110] was used for the generation of the other force field parameters of the solute;
  - the linker molecule was solvated in a buffer of chloroform by using the Amber module LEaP[98]. The solvation in chloroform was performed by selecting the keyword CHCL3BOX during the solvation process.
  - the solvated system was minimised for 100 cycles by using the steepest descend method and equilibrated at 300 K and 1 atm pressure for 10^5 MD steps with 2 fs time step (200ps) using the Amber module

Sander. During the equilibration the linker molecule was restrained to its starting position by using harmonic potential with a force constant of 10 kcal mol^{-1} $Å^{-2}$ and constraining bonds involving hydrogen atoms to their equilibrium distance.

- CM5 protocol
  - The linker molecule was geometrically optimized by using Gaussian 09[111] and selecting the basis set M06-L/MG3S. The calculation was performed until stationary point on the potential surface was found using the eigenvalue-following algorithm;
  - the Hirshfed population charge was calculated by using Gaussian 09[111] in the basis set M06-L/6-31+G(d,p) selecting as solvent environment the chloroform (Gaussian keyword scrf=(solvent=chloroform));
  - the Hirshfed charge were corrected in CM5 charges by using the CM5PAC software package[112];
  - GAFF[97;110] was used for the generation of the other force field parameters;
  - the molecule linker was solvated in a buffer of chloroform by using the Amber module LEaP[98]. The solvation in chloroform was performed by selecting the keyword CHCL3BOX during the solvation process.
  - the solvated system was minimised for 100 cycles by using the steepest descend method and equilibrated at 300 K and 1 atm pressure for 10^5 MD steps with 2 fs time step (200ps) using the amber module Sander. During the equilibration the molecule linker was restrained to its starting position using harmonic potential with a force constant of 10 kcal mol^{-1} $Å^{-2}$ and constraining the hydrogen bonds to their equilibrium distance.

Figure 3.7 shows the assembled system.

The previous two generated systems were simulated by using the MD module of the Sire-OpenMM software package[74]. Sire-OpenMM was in part detailed

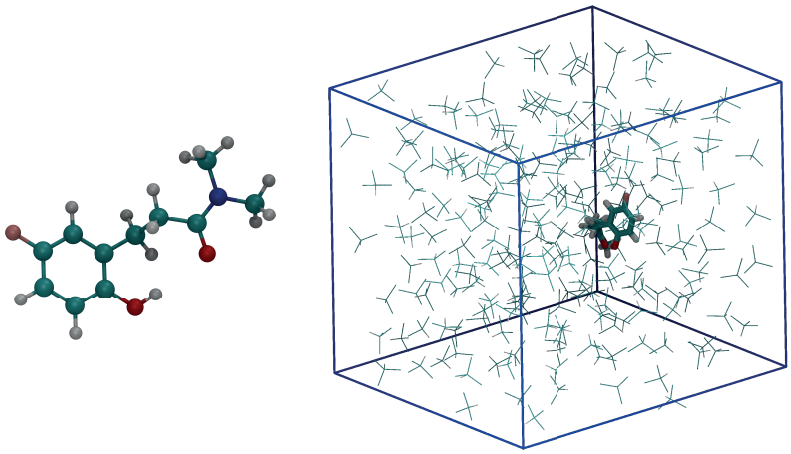


Figure 3.7: *The molecular system generated for the linker molecule with three carbon bonds along the alkyl carbon chain.*

in the second chapter for the calculation of the relative free energy of binding. However, this piece of software was also extended to perform fast MD simulation on GPUs by using the OpenMM APIs[75]. In particular, the mechanical potential 1.21 detailed in Chapter one was implemented, adding support for the long range electrostatic correction by using a reaction field[90;91]. The MD simulations were performed for 100 ns in the NPT ensemble setting the pressure and the temperature respectively to 1 atm and 300 K. The pressure was regulated by using MonteCarlo barostat[101;102] with isotropic scaling and an update frequency of 25 steps. The Andersen thermostat[82] was used to keep the temperature constant selecting a coupling coefficient of 10 ps^{-1}. The simulations were carried out by using the Leap-Frog Verlet integrator with a 2 fs time step. All the bonds were constrained to their equilibrium distances and the PME[113] cutoff scheme was used selecting a cutoff-distance of 10 Å and a tolerance error of 10^{-4}. During the simulations, the intra-molecular distance between the hydrogen of the OH in the 4-fluorophenol group and the oxygen atom in the amide group was monitored (Figure 3.8) and, the intra-molecular complexation energy was calculated by using equation 3.2. The histograms related to the intra-molecular distances by using the two previous protocols for the charge calculations and related to the linker molecule with thee carbon bonds are reported in Figure 3.9 along with the cal-

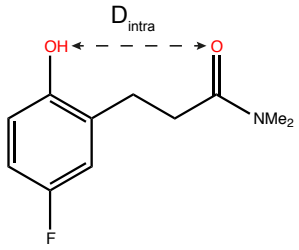


Figure 3.8: *In order to calculate the intra-molecular complexation energy, the intra-molecular distance D_{intra} was monitored along the MD simulations.*

culated intra-molecular complexation energies. The results have shown that the calculated binding energy are very different. The AM1-BCC method produces a free energy change of -4.35 kcal/mol preferring the folded state. On the other hand, the CM5 charge method produces a binding energy of +1.00 kcal/mol preferring the unfolded state. As previously described, the AM1-BCC method has been optimised for polar media simulations where TIP3P, SPC and TIP4P water model are frequently used. The CM5 method is a more robust charge method for force field parametrisation and the defined CM5 protocol allows the selection of the chloroform as solvent environment. Overall, the agreement between experimental and predicted data was considered closer by using the CM5 methods and, therefore, in this research project this protocol has been selected as standard protocol for the charge calculations and the linker molecule setting. The described CM5 protocol used to set the linker molecule system and the MD setting will be extensively used in the rest of this chapter with small changes where stated. As a consequence, this two protocols will be often referenced as the CM5 protocol and the MD settings or protocol.

In the CM5 protocol the linker molecule was solvated in a buffer of chloroform by using the Amber module LEaP and, specifying, the keyword CHCL3BOX [114] during the solvation process. A simulation test was conducted to highlight if the predefined chloroform atomic charges used by LEaP could significantly change the simulations. With this aim, a chloroform molecule was sketched by using Schrödinger Maestro [109] and the atomic charges were calculated as described in

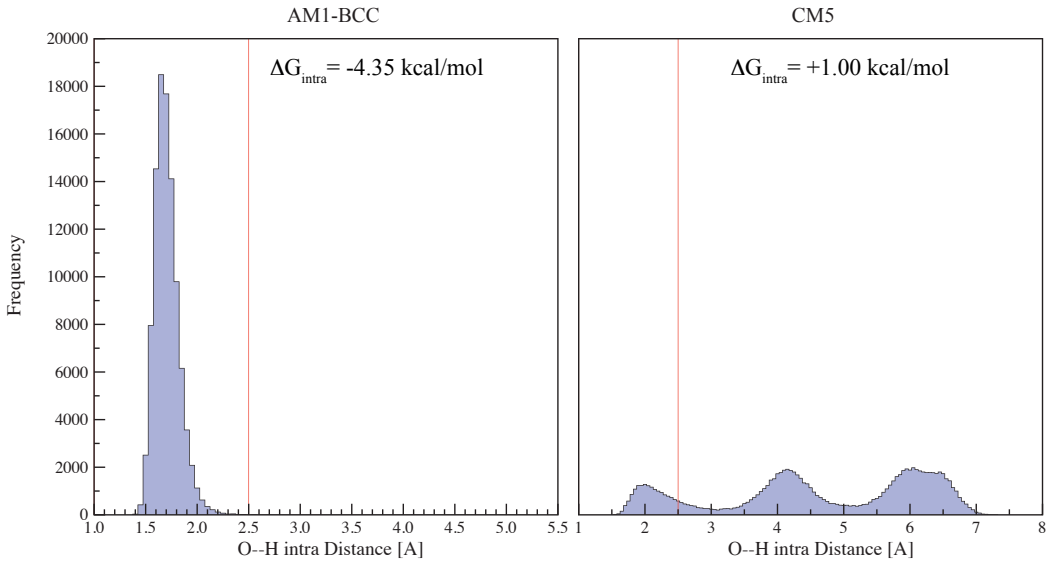


Figure 3.9: *The intra-hydrogen bond formation was monitored recording the intra-molecular distance D_{intra} shown in Figure 3.8 along a MD simulation. The threshold to have hydrogen bond formation was set to 2.5 Å (red vertical lines). The distance histograms obtained by using the AM1-BCC and the CM5 protocols are shown for the linker molecule with three carbon bonds along the alkyl chain. The intra-molecular complexation energy for both methods was calculated by using the equation 3.2. The probability of the folded state was estimated counting the frequency to observe the monitored intra-molecular distance to be less or equal to the selected threshold distance ($D_{intra} \leqslant 2.5$ Å).*

the CM5 protocol. Subsequently, the linker molecule with three carbon bonds along the chain was solvated in a box of chloroform by using the new chloroform molecule, without using the predefined LEaP chloroform model. A new MD simulation was conducted on the assembled system with the MD setting previously detailed. A comparison between the intra-molecular distance histograms produced by using the predefined chloroform model and the optimised version are shown in Figure 3.10 along with the calculated intra-molecular complexation energies. Significant differences were not observed and the standard LEaP chloroform model was used to simulate the other linker molecules.

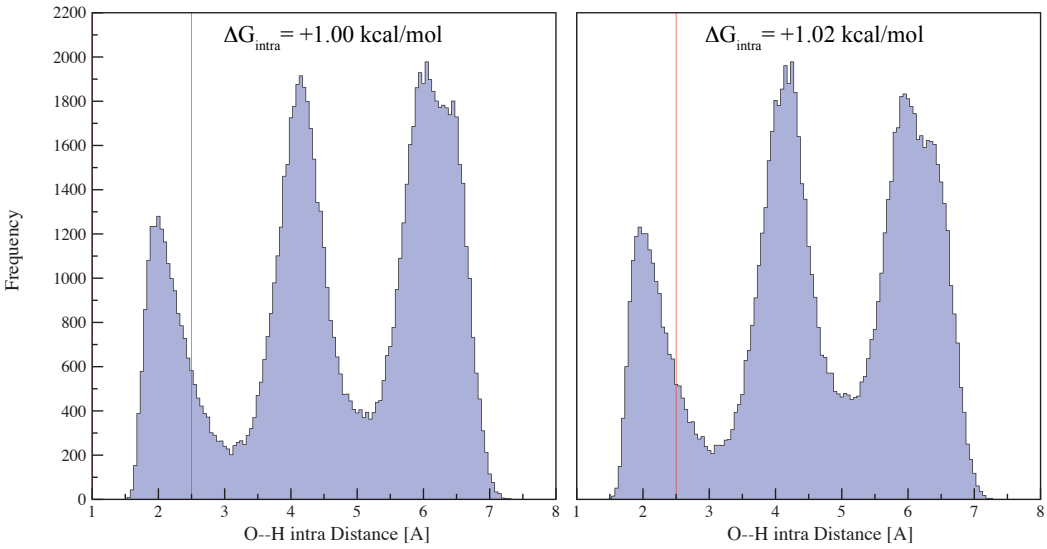


Figure 3.10: *In the CM5 protocol the solvation stage was executed by using a predefined model for the chloroform molecule. An optimised version of this molecule was modelled by using the CM5 atomic charge calculation. The intra-molecular distance histograms related to the linker molecule with three carbon bonds show that there are no significant differences between the optimised version (right histogram) and the standard parametrisation used by LEaP (left histogram).*

The CM5 and MD protocols previously used to set and simulate the linker molecule with three carbon bonds were also used to set and simulate all the linker molecules shown in Figure 3.2. Each simulation was repeated three times and the uncertainties were calculated as standard deviation of the mean over the three independent runs. For each system, the intra-molecular complexation

energy was calculated by using the equation 3.2 and results are reported in Figure 3.11 (a) along with the experimental data. In the comparison analysis between

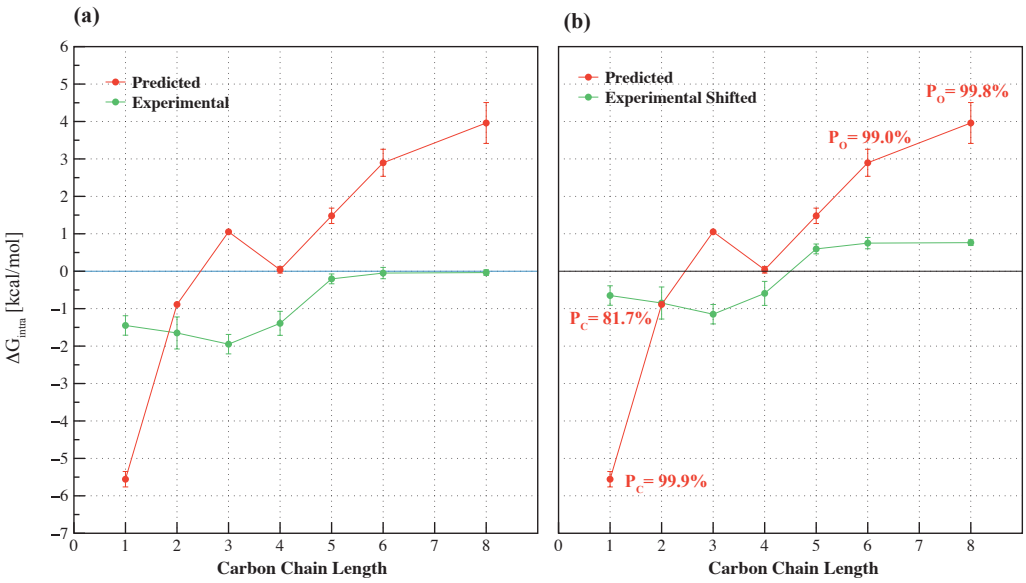


Figure 3.11: *(a) A comparison between the experimental and predicted intra-molecular complexation energies for the different linker molecules illustrated in Figure 3.2. (b) Experimentally the probability to observe 50% the folded state and 50% the unfolded state is measured between four and five carbon chain lengths. As a consequence, a correct comparison between the experimental and predicted data should shift the experimental measurements to have zero intra-molecular complexation energy when the carbon chain length is 4.5 carbon bonds.*

experimental and predicted data, a problem to face is the measured zero free energy point in the experimental data. Indeed, experimentally the probability to observe 50% the folded state and 50% the unfolded state was measured when the carbon length l is between $l = 4$ and $l = 5$ approximately. Therefore, a correct comparison should shift the experimental data to have zero binding affinity when the carbon length $l = 4.5$. To this end, the experimental intra-molecular complexation energy of the system with $l = 4$ and $l = 5$ were averaged and the resulting value of +0.8 kcal/mol was used to shift all the experimental intra-molecular complexation energies. Figure 3.11 (b) reports the shifted values. The predicted results validate the starting hypothesis that the increase in the carbon

chain length decreases the intra-molecular complexation energy. This is due to an increasingly favourable entropic contribution owing to the increased number of conformations available to longer linker groups in the "unfolded" state. However, there are discrepancies between the predicted and measured data especially for the extreme values of the carbon chain length. A plausible explanation could be the different "sensitivity" between the experimental measurements and the computational model. In Figure 3.11 (b) the experimental measurements show a plateau when the chain length $l \geqslant 5$. On the other hand, the computational model is significantly variable in this area. The unfolded population probability P_o have a difference of 0.8% between the systems with $l = 6$ and $l = 8$ however, this small variation in the probability population produces a change in the calculated intra-complexation energy of 1 kcal/mol which is significant in this context. In a similar way, the experimental intra-complexation energy between $l = 1$ and $l = 2$ does not significantly change while, the predicted value change is nearly 4 kcal/mol. In this case, the folded population probability P_c presents a difference of nearly 18%. To summarise, it should be possible to state that the experimental measurements at the extremes of the carbon chain length are not able to record small population changes between the unfolded and folded states; probably because they are lost within the experimental noise. It is also interesting to observe the presence of a local minima in the experimental and predicted data. However, the experimental minima is recorded when the carbon chain length $l = 3$ while, it is predicted when $l = 4$.

In order to improve the agreement between experimental and predicted data the LJ parameters related to the linker molecule systems assembled by using the CM5 protocol were modified to asses their impact on the simulations. In the MD module of Sire-OpenMM the LJ potential was implemented by using the equation 1.17 described in Chapter one. In this equation, the well depth parameter ϵ_{ij} depends on the atom i and j and it is computed by using the mixing rule:

$$\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j} \ , \tag{3.6}$$

where ϵ_i is a defined per atom force field parameter. In this case, the ϵ_i parameter of each linker molecule with two, three, four and five carbon bonds were multiplied by the factors 0.8 and 1.2. Each system was then simulated in a single run by using the MD setting previously detailed and the intra-comlexation energy was calculated. Results are reported in Figure 3.12. The impact on the simulation was not significant and no further investigations were performed.

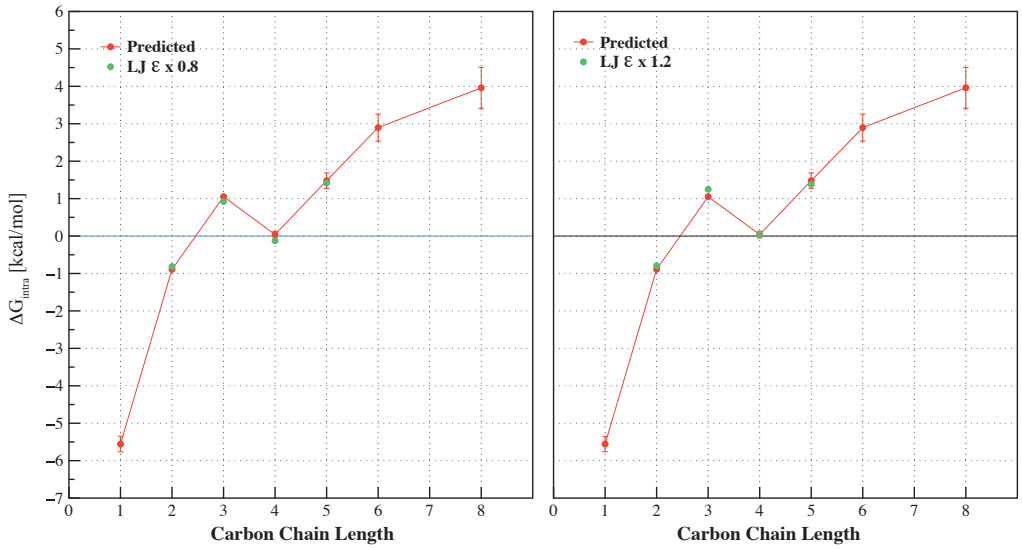


Figure 3.12: *The per-atom well depth parameter ϵ_i was multipled by the factors 0.8 and 1.2 for the linker molecules with two, three, four and five carbon bonds to check their impact on the calculated ΔG_{intra} values. The change in the intra-complexation energy is shown along with the previously predicted. No significant discrepancies were found and no further investigations were conducted.*

Another important factor that could affect the simulations is the carbon chain flexibility. This is mainly controlled by the force field parameters related to the dihedral angles along the carbon chain. In the MD module of Sire-OpenMM, the implemented dihedral potential energy is described by equation 1.20. The dihedral flexibility is mainly regulated by the amplitude parameter A_n in the cosine expansion. With the aim of improving the agreement between experimental and predicted data, the amplitude parameter A_n of each dihedral angles along the alkyl carbon chain was multiplied by the factors 0.3 and 0.8 for the system

with two, four and six carbon bonds. Again, each system was then simulated in a single run by using the MD protocol previously detailed and the intra-comlexation energies were calculated; results are reported in Figure 3.13. In addition different starting condition were tested with the chain starting in the folded and unfolded states but no significant differences were recorded (data not shown). The results highlighted that the simulations are not considerably affected by this change and no further investigation were performed.

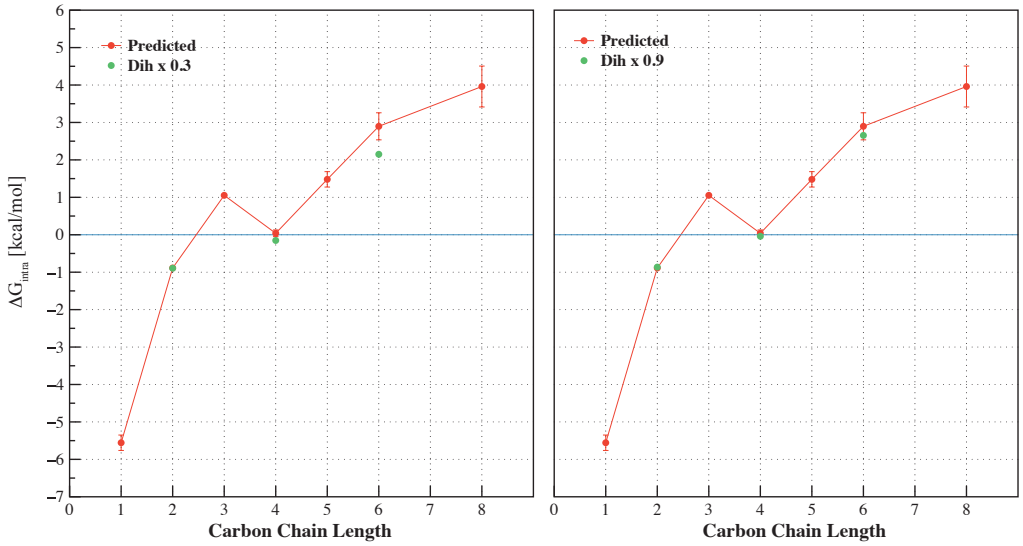


Figure 3.13: *The amplitude parameter of each dihedral angles along the alkyl carbon chain was multipled by the factors 0.3 and 0.8 for the linker molecules with two, four and five carbon bonds to check their impact on the calculated* ΔG_{intra}*. The change in the intra-complexation energy is shown along with the previously predicted. No significant discrepancy were found and no further investigations were conducted.*

In the experimental setup the different linker molecules were solvated in a solution of chloroform and tributylphosphine oxide in selected concentrations. As a consequence, in solution is present another molecule, the tributylphosphine oxide, which has not been considered so far in the molecular simulations. The reasoning behind this experimental hypothesis was that the thermodynamic process described in Figure 3.6 was not significantly affected by the tributylphosphine oxide molecule. Actually, the whole thermodynamic process is instead described as

illustrated in Figure 3.14. In this process there are four thermodynamic states in

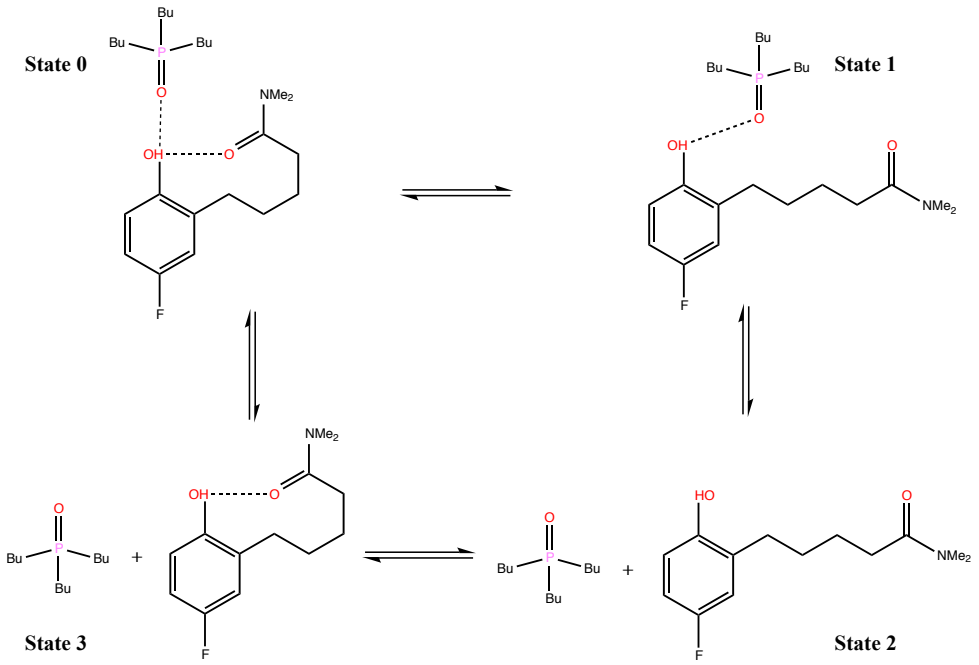


Figure 3.14: *In the experimental setup the full thermodynamic process can be described by using four thermodynamic states. The states 2 and 3 are respectively the "unfolded" and "folded" states previously introduced. In the state 0 the linker molecule can form intra-molecular hydrogen bond and inter-molecular hydrogen bond with the tributylphosphine oxide molecule while in the state 1, just the inter-molecular hydrogen bond is formed. One of the experimental hypothesis was that the state 0 was unlikely and, therefore, the tributylphosphine oxide molecule was ignored in the simulation setup considered so far.*

dynamic equilibrium, the "folded" and "unfolded" state previously introduced and two other states where the linker molecule is able to form or not inter-molecular hydrogen bonds with the tributylphosphine oxide molecule. For future references, these four states were named state 0, state 1, state 2 and state 4 as reported in Figure 3.14. The experimental hypothesis i.e. the fact that the tributylphosphine oxide molecule is not considerably affecting the thermodynamic process in Figure 3.6 and, therefore, the correctness of the computational models tested so far, could be proved or disproved if the state 0 is an unlikely state. In order to test this hypothesis, a model of the tributylphosphine oxide molecule was sketched by

using Schrödinger Maestro. Subsequently, its atomic charges were calculated by using the CM5 charge method and the molecule was assembled with the linker molecules in chloroform. Figure 3.15 illustrates an example of the produced system. In order to check the effect of the tributylphosphine oxide molecule in the

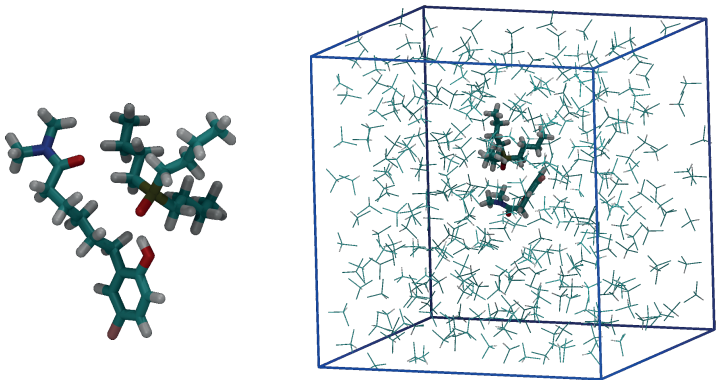


Figure 3.15: *The tributylphosphine oxide molecule was modelled by using the CM5 protocol and assembled with the linker molecules. The systems were also solvated in a box of chloroform.*

system simulations the linker molecules with two and six carbon bonds along the chain were simulated by using the MD protocol previously introduced. The intra- and inter- molecular distances reported in Figure 3.16 were recorded along the MD simulations and were used to discriminate between the four thermodynamic states as follows:

- state 0 : $D_{intra} \leqslant 2.5$ Å and $D_{inter} \leqslant 2.5$ Å
- state 1 : $D_{intra} > 2.5$ Å and $D_{inter} \leqslant 2.5$ Å
- state 2 : $D_{intra} > 2.5$ Å and $D_{inter} > 2.5$ Å
- state 3 : $D_{intra} \leqslant 2.5$ Å and $D_{inter} > 2.5$ Å

The MD simulations were repeated three times and the uncertainties were calculated as standard error of the mean over the three independent runs. Figure 3.17 reports the results for the simulated systems. In addition, Figure 3.18 shows an example of the intra- and inter- molecular distances recorded during a selected

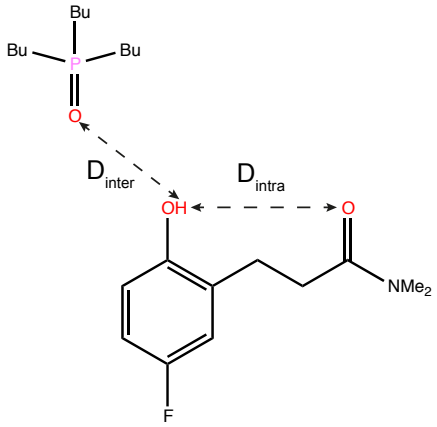


Figure 3.16: *The intra- and inter-molecular distances monitored along the MD trajectories.*

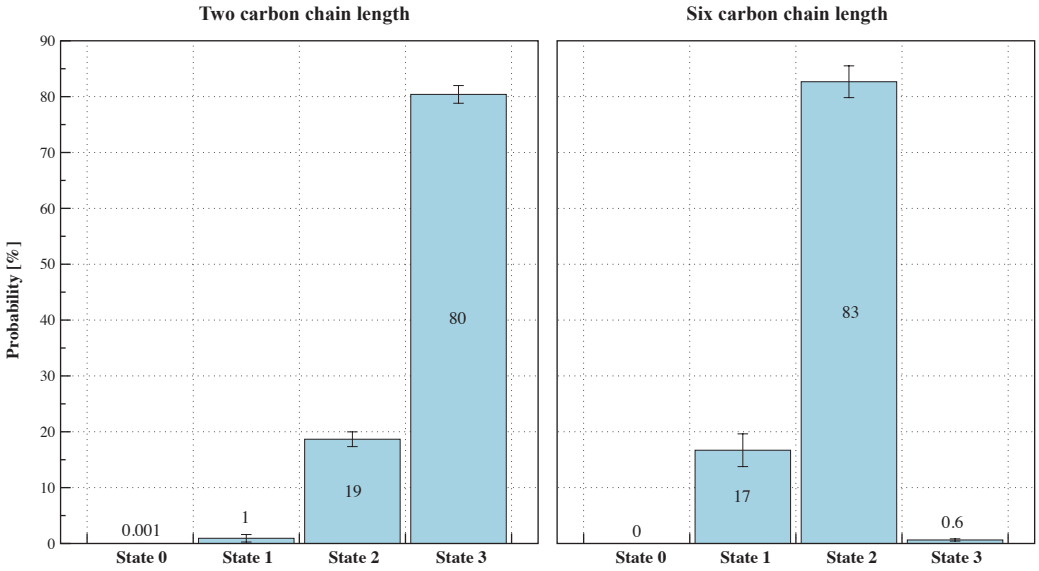


Figure 3.17: *The probability to find the system in the four thermodynamic states. The system with two carbon bonds along the chain (left figure) was mainly in the folded state 3 forming intra-molecular hydrogen bond. On the other hand, the system with six carbon bonds (right figure) preferred the unfolded state 2. In both simulations the state 0 was unlikely.*

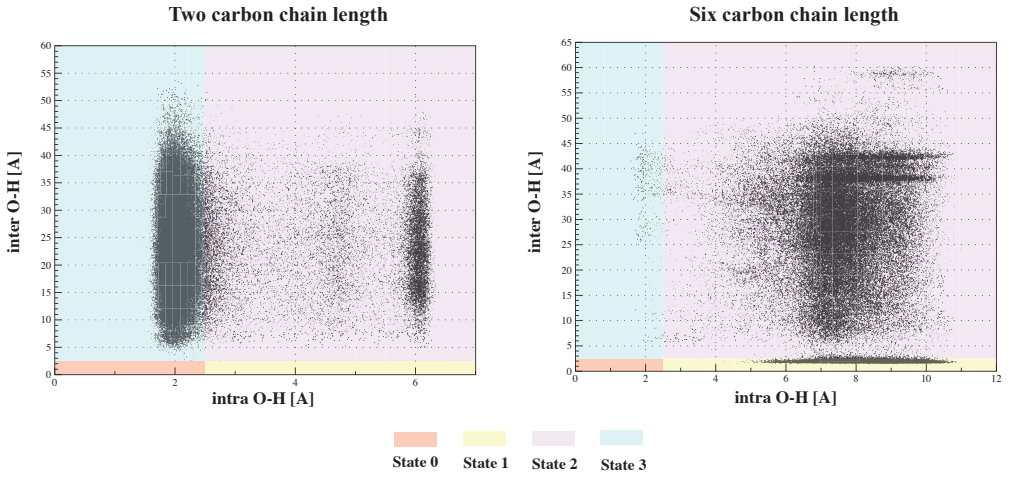


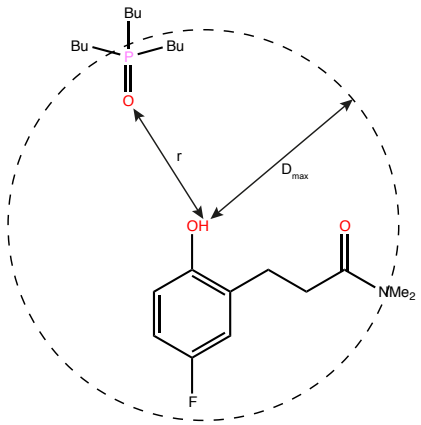
Figure 3.18: *The intra- and inter- molecular distances recorded during a selected MD run for the systems with two and six carbon bonds. The monitored distances are reported in Figure 3.16. The light coloured areas are the different four thermodynamic states while, the black shaded points are the recorded distances. The system with two carbon bonds (left figure) preferred the state 3 while the system with six carbon bonds (right figure) preferred the state 2.*

run. Figure 3.17 seems to confirm the experimental hypothesis that the state 0 was unlikely.

The previous results could be however biased by a tributylphosphine oxide concentration artefact. Indeed, the diffusive motions of the linker molecules and the tributylphosphine oxide molecule during the MD simulations could limit the binding event in the considered simulation time scale. In order to simulate the tributylphosphine oxide molecule in higher concentrations and, therefore, limiting the diffusive motion in solution, a restraint distance was implemented. The restraint distance was applied between the hydrogen atom of OH in the 4-fluorophenol group and the oxygen atom in the tributylphosphine oxide molecule. A special custom potential was implemented in Sire-OpenMM and applied to the selected atoms. The designed potential energy function $\mathcal{U}_{OH}$ used is:

$$\mathcal{U}_{OH} = \mathcal{U}_{SOH} + \theta(r - D_{max})(r - D_{max})^2 \ , \tag{3.7}$$

where $\mathcal{U}_{SOH}$ is the standard mechanical potential applied between the selected



$$\mathcal{U}_{OH} = \mathcal{U}_{SOH} + \theta(r - D_{max})(r - D_{max})^2$$

Figure 3.19: *The potential energy function used to constraint the tributylphosphine oxide molecule in a given sphere. The sphere has a radius D_{max} and it is centred on the hydrogen of OH in the 4-fluorophenol group. The potential was implemented to simulate the different tributylphosphine oxide concentrations in solution.*

atoms, D_{max} a selected distance threshold, θ the Heaviside function and r the distance between the selected atoms. Figure 3.19 illustrates these parameters. When the inter-atomic distance between the selected atoms is less than D_{max}, the implemented potential equates the standard mechanical potential otherwise, an harmonic contribution is added. As a consequence, the tributylphosphine oxide molecule is approximately enclosed in a sphere of radius D_{max} and, in the implemented case centred on the hydrogen of OH in the 4-fluorophenol group. Molecular simulations were performed by using the MD protocol and adding the defined $\mathcal{U}_{OH}$ potential. The systems with two and six carbon bonds with the tributylphosphine oxide molecule were simulated using different values of the parameter D_{max}: 4 Å, 8 Å and 12 Å. For each of this value, the simulation was repeated three times and the uncertainties were calculated as standard deviation of the mean over the three independent runs. Results are reported in Figures 3.20 and 3.21. In addition the system with four and five carbon bonds were simulated

with the same setting but in a single run without repetitions (data not shown).

The state 0 was unlikely to happen for all the simulated systems and with the range of considered tributylphosphine oxide concentrations. Therefore, the tributylphosphine oxide and the linker molecules do not bind at the same time. This result proved the correctness of the experimental hypothesis and the best agreement between the predicted and experimental data reached in this research project is shown in Figure 3.11.

3.4 Chapter Conclusions

This chapter investigated the effect of molecular flexibility on conformational equilibria of a set of molecules with two main moieties linked together by flexible carbon chains. The molecules can experimentally adopt different conformers in chloroform solution and, in particular, two thermodynamic conformations were defined the "folded" and "unfolded" states. In the "folded" conformation an intra-molecular hydrogen bond was formed and an in the "unfolded" state the hydrogen bond was broken. The intra-molecular complexation energy between these two states was computed by using molecular dynamics simulations and compared to the experimental data. The agreement was not optimal but in many cases it was possible to find reasonable explanations. With the aim of improving the agreement, force field parameters were changed. The partial charges were computed by using the AM1-BCC and the CM5 methods. The latter approach was preferred because of closer results to the experimental values. Other force field parameters were changes such as LJ and dihedral parameters but not significant changes were observed. The partial charge seems to have the major impact in the calculations. It would be possible to improve the agreement, for instance studying the effects of the CM5 charge scaling or by using more advanced force fields such as polarizable force fields or QM/MM methods.

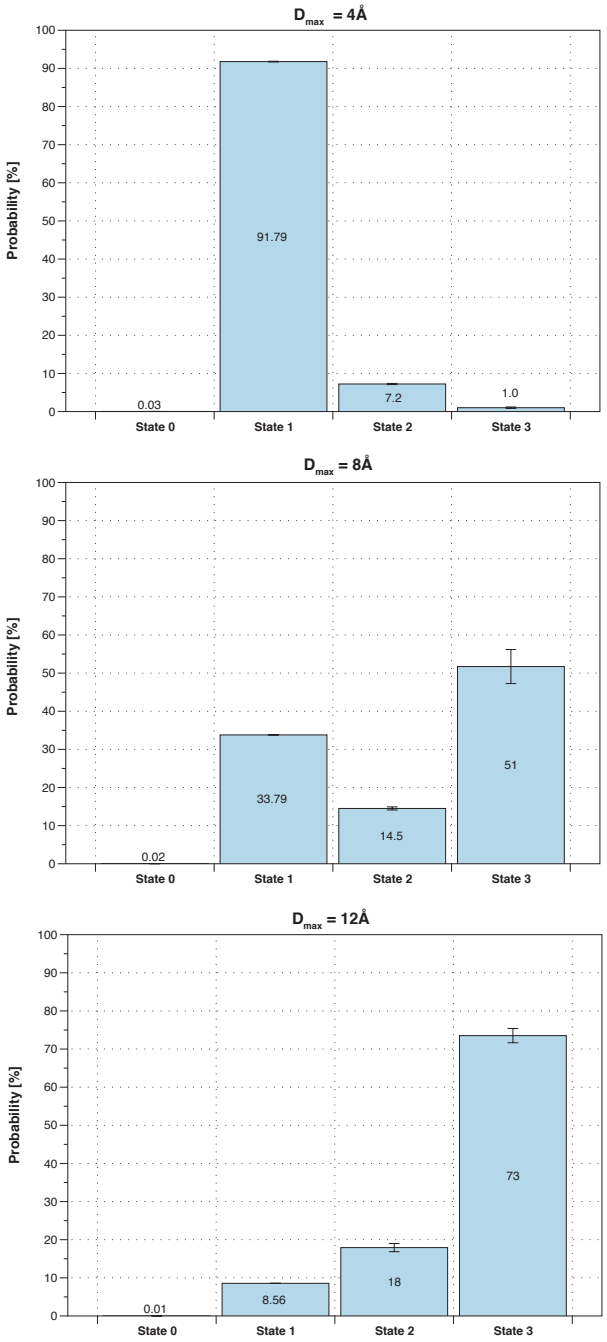


Figure 3.20: *The probability to observe the four thermodynamic states for the linker molecule with two carbon bonds and for different values of D_{max} (equation 3.7) parameter is shown. This parameter is used to simulate the tributylphosphine oxide concentrations. Low values of D_{max} simulate high concentrations. The results show that if the simulated tributylphosphine oxide concentration is in the selected range D_{max} = 4 Å, the tributylphosphine oxide binds the linker molecule. On the other hand, if the concentration is D_{max} = 12 Å the intra-molecular folded state is preferred. The state 0 was unlikely for all the simulated tributylphosphine oxide concentrations.*

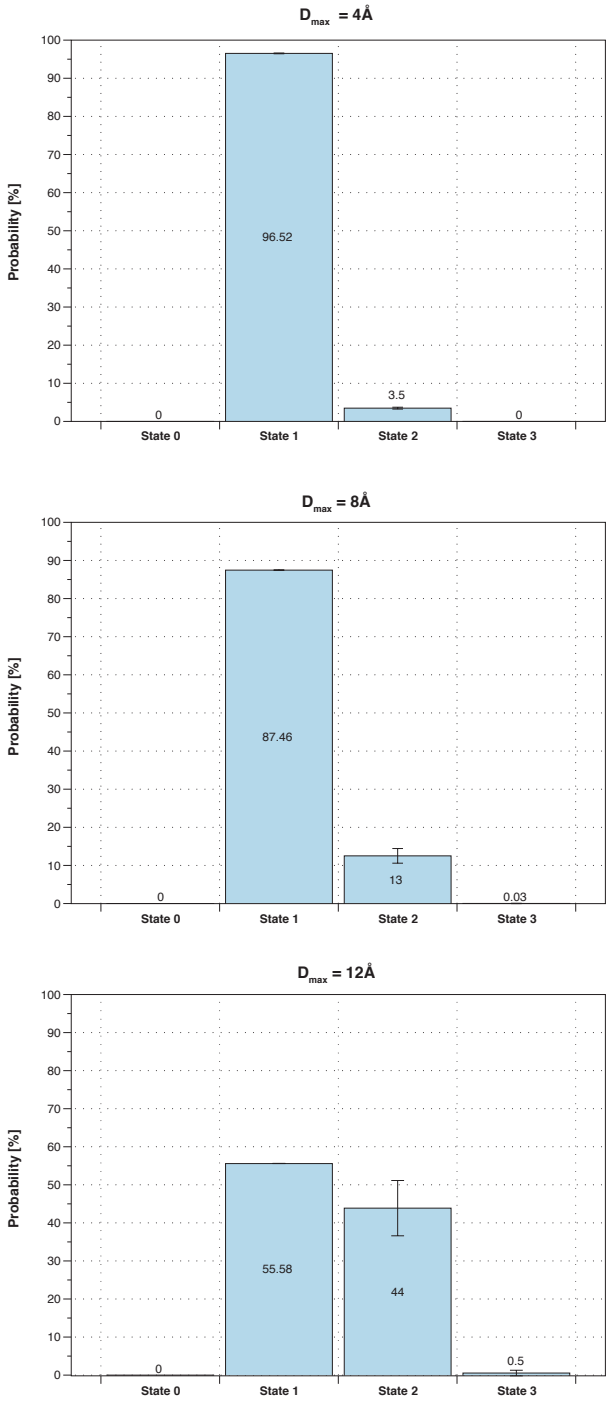
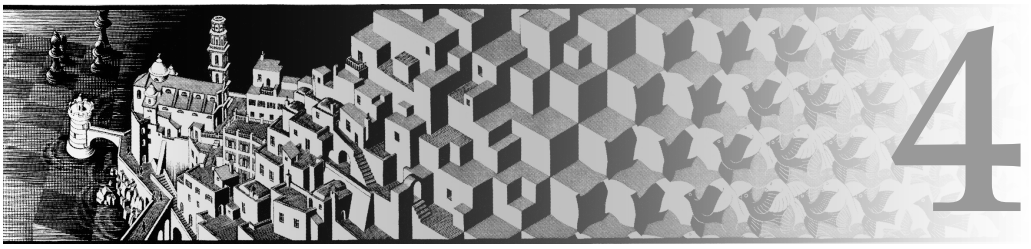


Figure 3.21: *The probability to observe the four thermodynamic states for the linker molecule with six carbon bonds and for different values of D_{max} (equation 3.7) parameter is shown. This parameter is used to simulate the tributylphosphine oxide oxide concentrations. Low values of D_{max} simulate high concentrations. The results show that if the simulated tributylphosphine oxide concentration is high in the selected range (D_{max} = 4 Å), the tributylphosphine oxide binds the linker molecule and this happened also for lower concentration values. Also in this case the state 0 was unlikely for all the simulated tributylphosphine oxide concentrations.*

"I still remember the day when my father gave me my first computer. I was only ten, and at the time it was just an extraordinary machine to play with. For me, the computer is still that remarkable toy that I had as a child, but now I have changed game. It is a fantastic box that can bring ideas and dreams to life, and to understand where all those interminable calculations can come to an end is like leaping ahead, through human thoughts"

— Gaetano Calabró



4 Non Additivity

4.1 Introduction

As shown in Chapter 1.2 a significant aspect along the critical path for drug discovery is the optimization of the non-covalent binding between a disease-involved protein and small organic molecules. To this end, free energy of binding is frequently used to quantify the protein-ligand interactive strength and iterative SAR studies are performed on promising hits to improve the binding affinity. Unfortunately, nowadays, the SAR stage seems to be more an art to master than a standardised protocol to be routinely applied by medicinal chemists. The main difficulties are correlated to the optimisation of enthalpy and entropy components. Binding enthalpy changes are notoriously difficult to improve, and they are related to two concomitant effects: the VdW forces/hydrogen bonds and desolvation of polar groups[13]. VdW forces are optimised by the perfect shape complementarity between the biomolecular target and

the drug-like molecule, while hydrogen bonds are maximised when the distances and angles between hydrogen bond donors and acceptors are optimal in the complex[13]. The desolvation of polar groups reflects the strength of the interactions between the solvent-target and solvent-ligand before the formation of the complex. A favourable enthalpy change is an indication of a sufficiently strong interaction between the target and the ligand that compensates for the unfavourable enthalpy change related to desolvation[13]. On the other hand, entropy changes are driven by two major terms: conformational entropy changes and solvation entropy. The former is related to the reduction of translational, rotational and internal degrees of freedom of ligand and protein after the binding while the latter depends on the release of water molecules from the binding site[13] and ligand in solution.

In this intricate framework, a complex task is to rationalise and simultaneously optimise all the different contributions at the atomic level and to find general rules to translate into protocols and guidelines. In recent years, meaningful improvements have been achieved in the structure determination of protein-ligand complexes using X-ray crystallography and NMR spectroscopy and, a plethora of computational methods based on static structures such as AutoDock Vina[115] or idock[116] have been applied to the prediction of the binding free energy. In order to find "hits" with sufficient affinity to be considered interesting, these methods usually need to screen thousand compounds per day and, therefore, the approximations introduced to achieve throughput cause the predicted binding free energy to be frequently inaccurate[117;118].

One of the main simplifications introduced by these approaches is the additivity of the binding free energy i.e. the assumption that free energy and free energy changes can be decomposed into sum of independent components ascribed to specific parts in a system. The extent to which this hypothesis can be considered valid is a question that is of central importance in many chemical and biochemical contexts. In general, if the cruel question is: "Can the free energy or entropy and their changes be decomposed into a sum of independent components ascribed to specific parts in a system", then the answer is negative. Free energy and entropy are indeed global properties of the whole phase space and

their component decompositions hold in the hypothesis that the phase space is divisible into uncorrelated parts. This result has been clearly proved in Statistical Mechanics[119]. In equation 1.14, the Helmholtz free energy is given by:

$$F = k_b T \ln\langle e^{\beta\mathcal{H}}\rangle_{NVT} \ .$$

A minimum requirement to express the total free energy as sum of components is to express the system Hamiltonian into sum of components e.g. two parts $\mathcal{H}_1$ and $\mathcal{H}_2$ related to two different system interactions. In this case it is always possible to state that the total system energy E is equal:

$$E = \langle\mathcal{H}\rangle = \langle\mathcal{H}_1 + \mathcal{H}_2\rangle = \langle\mathcal{H}_1\rangle + \langle\mathcal{H}_2\rangle \ . \tag{4.1}$$

However, this property is not transferred to the Helmholtz free energy:

$$F = k_b T \ln\langle e^{\beta\mathcal{H}_1} e^{\beta\mathcal{H}_2}\rangle_{NVT} \ , \tag{4.2}$$

where the terms in the ensemble average cannot be further factorized. However, if the system Hamiltonian $\mathcal{H}$ can be broken down into two components i.e.:

$$\mathcal{H}(q_1, q_2, p_1, p_2) = \mathcal{H}_1(q_1, p_1) + \mathcal{H}_2(q_2, p_2) \ , \tag{4.3}$$

where q_1, p_1, q_2 and p_2 are uncorrelated phase space coordinates, then the property holds:

$$F = k_b T \ln\langle e^{\beta\mathcal{H}_1}\rangle_{NVT} + k_b T\langle e^{\beta\mathcal{H}_2}\rangle_{NVT} = F_1 + F_2 \ . \tag{4.4}$$

Despite this result, the additivity of free energy has been observed in many chemical systems where covalent interactions are involved. In this cases the additivity assumption has been validated and it is used to predict chemical equilibria and kinetic[120]. However, in systems such as protein-ligand or protein-protein where the relevant interactions are non-covalent most of the time this hypothesis cannot be taken for granted. As a result, in drug design, the non-additivity or cooperativity could significantly affect the structural-activity stage. Indeed, two ligand

fragments linked together can result in a ligand with a binding affinity that could be greater (positive cooperativity) or lower (negative cooperativity) than the sum of its parts[76]. This fact has been conveniently ignored for many years in the protein-ligand binding context in favour of the simpler additivity model. Patel and co-workers[121] investigated 19 different biological systems to examine the extent of non-additive substituent effects on binding affinities and they found that only half exhibited additive behaviour.

An in depth study of non-additivity of functional groups in protein-ligand interactions was performed by Baum et al.[76] and is central to this thesis. In the case study the Thrombin protein and series of congeneric inhibitors were considered. Thrombin is a serine protease of the chymotrypsin family[122] involved in the hemostasis, the delicate balance between bleeding and thrombosis, which is essentially maintained by the localisation and rapid amplification of coagulation proteinases and their cofactor complexes at the site of vascular injury[123]. Thrombin once is generated in the blood from its inactive precursor prothrombin, plays two important and paradoxically opposing functions. It acts as a procoagulant factor when it converts fibrinogen into an insoluble fibrin clot that anchors platelets to the site of lesion and initiates processes of wound repair[124]. In this cascade prrocess thrombin cleaves fibrinogen to fibrin, activates the fibrin-cross-linking transglutaminase factor XIII (FXIII), catalyzes its own generation through activation of FXI, FVIII, and FV, and stimulates platelet aggregation via cleavage of the membrane-bound protease-activated receptors (PARs) 1, 3, and 4[122]. In contrast, thrombin acts as an anticoagulant through activation of protein C. This function unfolds in vivo upon binding to thrombomodulin, a receptor on the membrane of endothelial cells. Binding of thrombomodulin suppresses the ability of thrombin to cleave fibrinogen and PAR1, but enhances >1000-fold the specificity of the enzyme toward the zymogen protein C[124]. Hijacking of thrombin by thrombomodulin and activation of protein C in the microcirculation constitute the natural anticoagulant pathway that prevents massive intravascular conversion of fibrinogen into an insoluble clot upon thrombin generation[124;125;126].

There are several indications for an allosteric behaviour of thrombin. First,

prothrombin undergoes large conformational changes during activation[122]. Second, the Basic Pancreatic Trypsin Inhibitor (BPTI), in spite of its relatively bulky reactive site loop, can tightly bind to the thrombin Glu192 Gln mutant[122]. The structure of the BPTI complex with this thrombin mutant (1BTH) explained the improved binding through favourable interactions of the Gln 192 carboxamide group with BPTI but also revealed a dramatic opening of the active site cleft, allowing accommodation of the bulky inhibitor[122]. Third, Na+ has been found to be an important allosteric modulator of α-thrombin[122]. Kinetically, two allosteric states, a "slow" and a "fast" α-thrombin form, have been defined, which are characterised by the absence and presence, respectively, of an Na+ ion, bound in the range of the physiological Na+ concentration[122]. For example, the fast thrombin form cleaves fibrinogen as well as the protease-activated receptors (PARs) more efficiently, i.e., displays procoagulant, prothrombotic, and prosignaling properties, while the slow form preferentially cleaves protein C and thus exhibits more anticoagulant properties[122;127].

Many diseases including stroke and myocardial infarction involve thrombosis; therefore, thrombin is a preferred target of antithrombotic drugs[128]. Drugs available to block thrombin action include heparins, hirudins (lepirudin and bivalirudin), vitamin K antagonists and a new generation of direct thrombin inhibitors such as dabigatran and argatroban[128]. In addition, the association of idiopathic venous thrombosis with occult cancer is generally recognized. However, it has not been fully appreciated that thrombin generated during thrombosis can augment the malignant phenotype[129]. Indeed, Thrombin protein activates tumor cell adhesion to platelets, endothelial cells, and subendothelial matrix proteins, it enhances tumor cell growth and increases tumor cell seeding and spontaneous metastasis[129].

The complexity of thrombin function and regulation have captured the interest of many investigators over the years and therefore many studies have been conducted. Therefore, it is an easily accessible protein with well established and accurate crystallographic structure. A three dimensional representation of this protein with its binding site is reported in Figure 4.1. Structurally the protein

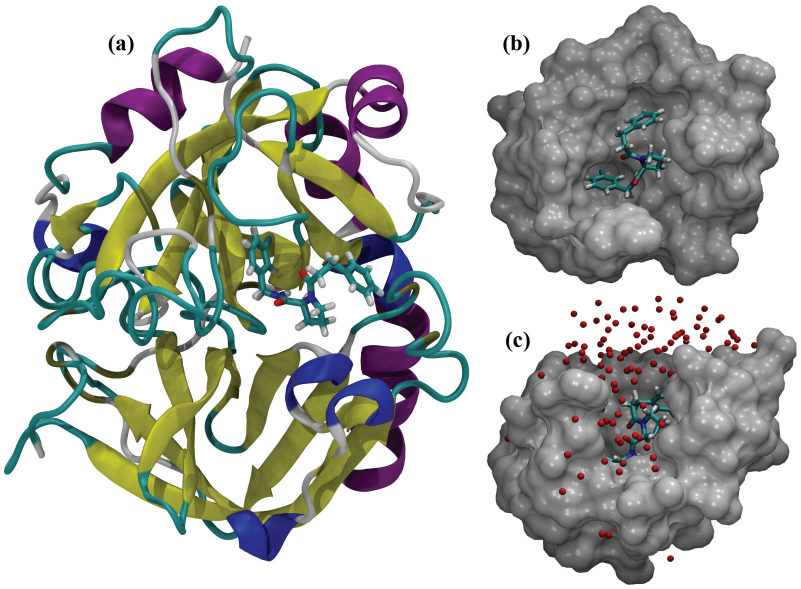


Figure 4.1: *(a) Tertiary structure of the modelled Thrombin protein and one of its inhibitors. (b) A representation of the binding site of Thrombin and (c) its exposition to water molecules represented as red spheres.*

presents three aligned sub-pockets illustrated in Figure 4.2 along with the main protein residues involved in the binding and shows the main protein inhibitors considered in this study.

As previously mentioned, the Baum and co-workers [76] investigation of Thrombin and its non-additivity with specific inhibitors was significant to this research project. In the investigation, the Isothermal Titration Calorimetry (ITC) was used to measure the free energy of binding of four series of congeneric inhibitors of Thrombin. The inhibitors differ in three main structural modifications. The first modification is the presence or absence of a terminal amino group ($X = H$, $X = NH_2$) that interacts with the carbonyl oxygen of Gly216, forming hydrogen bonds. A second modification is related to the group that secures the ligand in the S1 pocket. Either a meta-chlorobenzyl or para-benzamidine moiety is attached to the L-prolyl portion via an amide bond [76]. Finally, a third modification concerns the side chain R^2 anchored to the sub-pocket S3. A series of methyl to benzyl groups are added to fill up the S3 hydrophobic pocket. The study revealed a

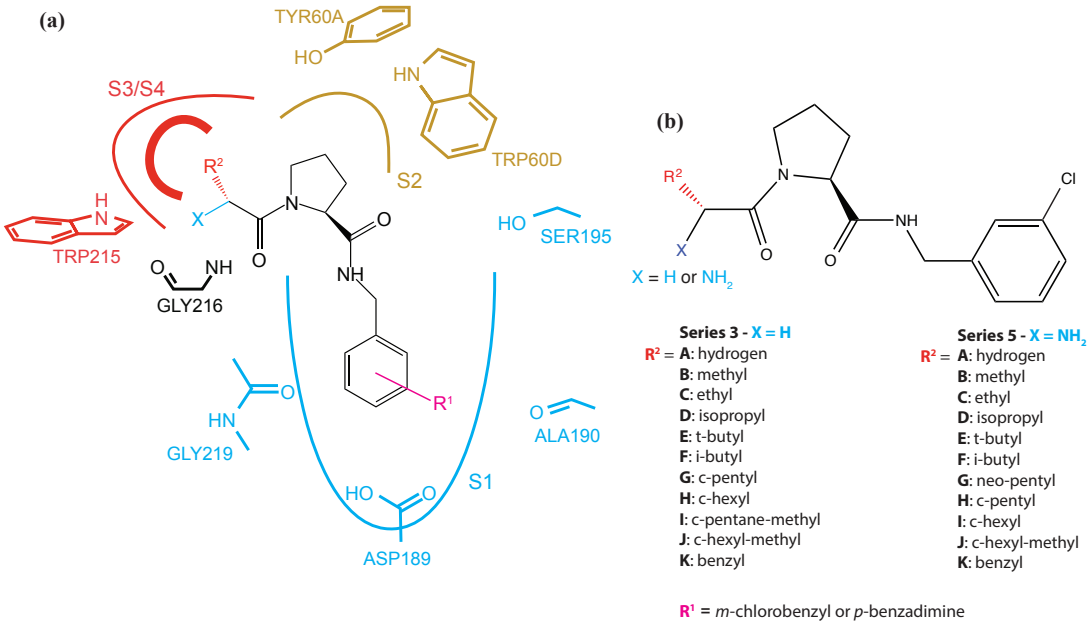


Figure 4.2: *(a) The Thrombin protein and the main residues involved in the binding with series of congeneric inhibitors. (b) The ligands present three main modifications at the position X (X = H or NH_2), the group R^2 anchored in the sub-pocket S3 and the group R^1 anchored to the sub-pocket S1. When the group R^1 is m-chlorobenzyl the modification related to the position X divides the inhibitors into two groups named **3** (X = H) and **5** ($X = NH_2$) series. Figure adapted from* Baum et al. *(76).*

positive linear correlation between the hydrophobic contact surface area of the R^2 groups with the protein and the free energy of binding (Figure 4.3). The increase in affinity is caused by the increase in size of the hydrophobic occupants of the S3 pocket. In addition, the presence of the amino group produces a further reduction of the binding free energy due to the additional hydrogen bond with the Gly216. However, the amino group also seems to interact cooperatively with the hydrophobic binding pocket S3. Assuming independence of the contributions added by the single interactions, the presence or absence of the amino group should produce a translation of the correlation line but not a slope change[76] (Figure 4.3). This

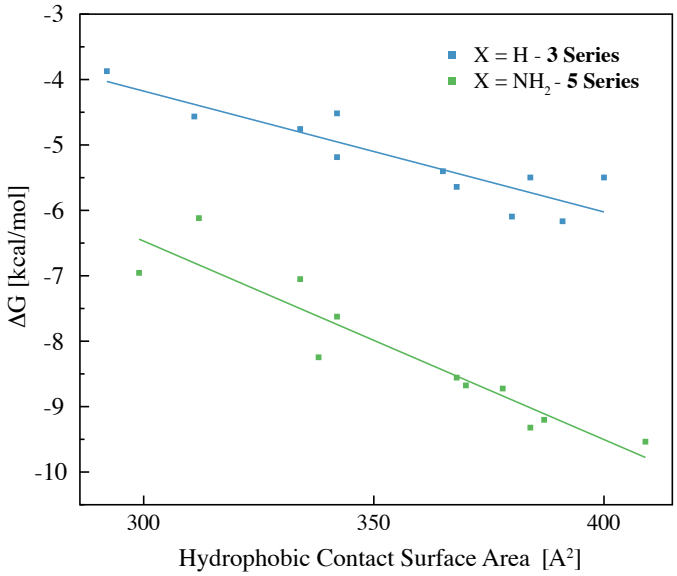


Figure 4.3: *The correlation of hydrophobic ligand surface area in contact with the protein is plotted against the Gibbs free energy of binding. The plot shows the ligands where the functional groups in the X and R^2 positions are changed and R^1 is the meta-clorobenzyl group. A linear correlation is obtained for the 3 ($X = H$) and 5 series ($X = NH_2$) but the different line slopes suggest the presence of non additivity in the system. Figure adapted from* Baum et al.[76]

suggests that there is cooperativity between the hydrophobic free energy component related to the sub-pocket S3 and the free energy component related to the hydrogen bond formation with the Gly216. The effect is also confirmed by the thermodynamic analysis. For example in Figure 4.4 the binding free energy change measured for the ligand 5K ($3B \rightarrow 5K$), which includes the modification

of both ligands 3K and 5B is -4.6 kcal/mol while, the single contributions are respectively for the ligands 3K ($3B \rightarrow 3K$) and 5B ($3B \rightarrow 5B$) -0.9 kcal/mol and -1.6 kcal/mol. For this ligand a binding free energy change of -2.5 kcal/mol would be expected if the interactions were additive; the affinity increases instead by -4.6 kcal/mol (Figure 4.4).

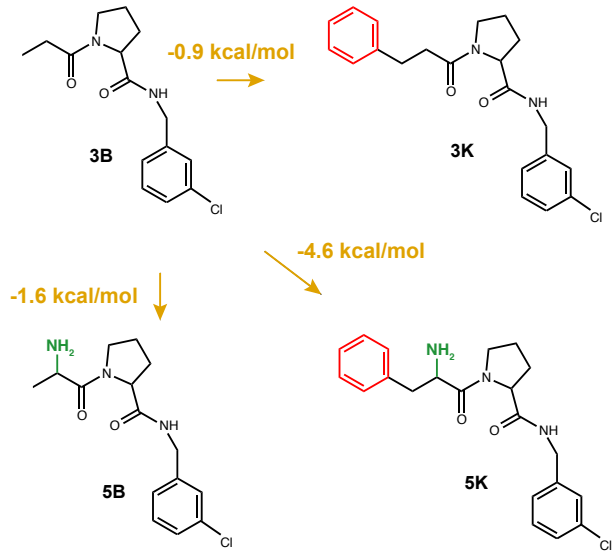


Figure 4.4: *Binding free energy differences $\Delta\Delta G$ between selected thrombin inhibitors. The compound 5K presents the modification of both the ligands 3K and 5B. However, the binding free energy differences $\Delta\Delta G(3B \rightarrow 5K) \neq \Delta\Delta G(3B \rightarrow 3K) + \Delta\Delta G(3B \rightarrow 5B)$. This is an experimental evidence that free energy is not an additive property.*

The previous case study is an example where the additivity assumption could have misdirected medicinal chemists in the lead optimization stage. In the Baum and coworkers[76] investigation a possible interpretation of the cooperativity effect was explained by the dynamic properties of the inhibitors bound to the active site. An analysis of the B-factor measured for the different thrombin inhibitors in complex was performed by using X-ray diffraction. This parameter reflects the average atomic fluctuations and is frequently used to gain information related to the mobility of residues in crystalline structures. Figure 5.3 illustrates B-factor for the 3L and 4L ligands. The colour scale changes from dark blue to red, indicating low and high values, respectively, of the B-factor[76]. The results show that the

presence or absence of the amino group produces a rigidification of the pocket S3 (dark blue in Figure 5.3). The entropy change ($-\Delta\Delta TS$) increases by +3.3 kcal/mol, which is partially balanced by a decrease in enthalpy change ($\Delta\Delta H$) of -5.3 kcal/mol.

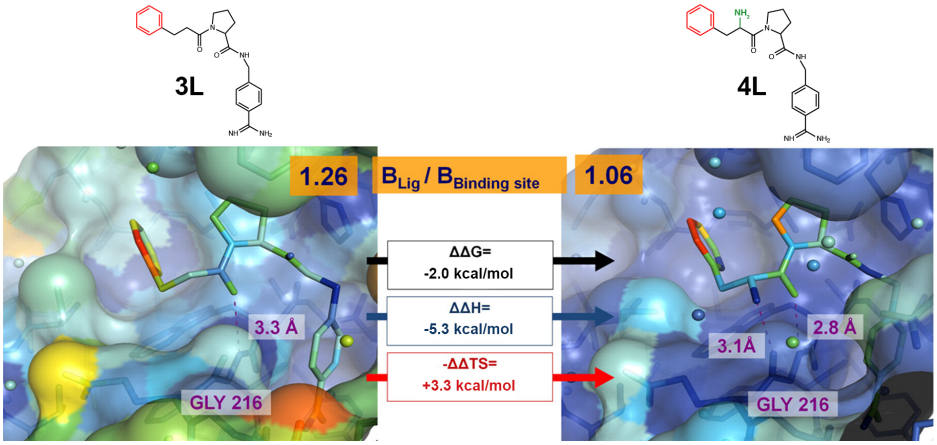


Figure 4.5: *The crystallographically determined binding mode of* 3L *and* 4L *in complex with thrombin. Colors are assigned to all atoms according to their temperature factors, from blue (low B-factor) to green to yellow and to red (high B-factor).The ratio of the mean B-factor of ligands with respect to the active site residues is indicated. The binding free energy change and its enthalpic and entropic components are also given. The ligand* 4L *forms a charge-assisted hydrogen bond (3.1Å) and a slight reduction of the adjacent hydrogen bond (distance 3.3Å vs 2.8Å) between the ligand carbonyl and the nitrogen of* Gly216 *is observed. A significant decrease in B-factor ratio between the ligand and the binding site from 1.26 to 1.06 was observed. (Figure adapted from* Baum and coworkers.[76]*)*

The previous analysis highlighted the importance of dynamics in the protein-ligand binding process to rationalise and explain the sources of non-additivity. Indeed, protein and ligand fluctuations around average structures are very important to capture entropic effects that cannot be accounted using solely rigid structures produced by crystallographic methods. Despite their computational cost compared to docking and scoring function methods, free energy calculations account for enthalpy and entropy changes in protein-ligand binding, usually producing improved results[117;130]. The final two chapters of this thesis had two main goals: reproduce in-silico the non additivity of binding free energy in the

Thrombin system obtained by Baum and coworkers[76], which is described in this chapter and, explain the possible non-additivity sources in the system, which is detailed in Chapter five. The investigation was performed by using the implemented free energy code described in Chapter two and based on the alchemical transformation method accelerated with the latest General Purpose Graphic Processing Unit (GPGPU) technology.

4.2 Thrombin Molecular Modelling and Setup

In this study the non-additivity of the Thrombin inhibitors in the 3 and 5 series was considered. In particular these series are obtained changing the hydrophobic group R^2 anchored to the sub-pocket S3, the hydrogen with an amino group and selecting the functional group R^1 to the *m*-chlorobenzyl in the sub-pocket S1 (Figure 4.2 (b)). In order to reproduce the non-additivity in-silico the crystallographic structure of human thrombin in a complex with a thrombin ligand structurally related to the ligands simulated in this study was downloaded from the PDB databank (PDB code $2ZC9$[76]). The protein was inspected and revised using Schrödinger Maestro[109]. The hirugen chain was removed from the structure. The side-chain of Arg75 in chain H was completed in a solvent exposed conformation. The incomplete light chain was capped before Glu1C with an ACE residue and after Ile14L with an NME residue. The incomplete heavy chain was capped after Gly246 with an NME residue. Missing residues Trp148, Thr149, Ala149A, Asn149B, Val149C, Gly149D, Lys149E in chain H were modelled in the structure using the FALC-Loop web server[131]. Standard protonation states were assumed for protein side-chains. On the basis of visual inspection of hydrogen bonding patterns, His57 and His71 were modelled in their uncharged, δ-tautomer. His91, His119 and His230 were modelled in the δ-tautomer[132]. Disulfide bridges were modelled between Cys42-Cys58, Cys1-Cys122, Cys168-Cys182 and Cys191-Cys220[132]. The ligands in the 3 and 5 series were modelled by using Maestro and manually placed inside the binding site in a starting conformation selected according to the crystallographic data and visual inspection. In addition, all the amino groups related to each ligand in the 5 series were protonated. In order

to assemble input files for all the different complexes the FE-Setup[96] software package was used. This piece of software is able to automatically create the solvated complexes and ligands starting from their PDB structures. In particular the following protocol was set in FE-Setup for the preparation of the ligands, protein and complexes:

- Ligands
  - The atomic point charges were assigned by using Antechamber[98] selecting the AM1-BCC[28] method;
  - GAFF[97;110] was used for the generation of the force field parameters;
  - the ligands were solvated in a buffer of water selecting TIP3P water model using the Amber module LEaP. In addition, for the ligands in the 5 series the solution was also neutralised adding Cl^- counter ions due to the positive net charge after the ligand protonation;
  - the solvated systems were minimised for 100 cycles by using the steepest descent method and equilibrated at 300 K and 1 atm pressure for 10^5 MD steps with 2 fs time step (200 ps) using the Amber module Sander. During the equilibration stage, the ligands were also restrained to their starting positions using a harmonic potential with a force constant of 10 kcal mol^{-1} $Å^{-2}$ and constraining the hydrogen bonds to their equilibrium distances.
- Protein
  - Amber ff99SB[133] force field parameters were used to parametrise the protein;
  - the protein was minimized in vacuum for 500 cycles of steepest descent method using the Amber module Sander.
- Complexes
  - The ligands were combined with the Thrombin protein and solvated in a buffer of water selecting TIP3P water model by using the software LEaP. Counter ions were also added to neutralise the solution;

– the complexes were minimised for 500 cycles by using the steepest descent method and equilibrated at 300 K and 1 atm pressure for 10^5 MD steps with 2 fs time step (200 ps) using the Amber module Sander. During the equilibration stage, the protein and the ligands were restrained to their starting positions using a harmonic potential with a force constant of 10 kcal mol^{-1} Å^{-2} and constraining the hydrogen bonds to their equilibrium distances.

In order to quantify the Non Additivity (NA) level for each ligand presents in the system it is convenient to consider the Figure 4.6. The generic ligand $5X''$ presents the functional group R'' and the amino group NH_2 compared to the generic base scaffold $3X'$. Therefore, the ligand exhibits both the modifications of the ligands $3X''$ and $5X'$. As a consequence, the level of non-additivity can be defined as:

$$NA = \Delta\Delta G_{3X'\rightarrow 5X''} - (\Delta\Delta G_{3X'\rightarrow 5X'} + \Delta\Delta G_{3X'\rightarrow 3X''}) \ . \tag{4.5}$$

The thermodynamic equation can be rewritten by using the thermodynamic cycle with the end states $3X'$, $5X'$ and $5X''$:

$$NA = \Delta\Delta G_{5X'\rightarrow 5X''} - \Delta\Delta G_{3X'\rightarrow 3X''} \ . \tag{4.6}$$

If NA equates zero then the system is additive, otherwise it quantifies the non-additivity level.

To compute non-additivity, the relative binding free energy between the different ligands in the two series was calculated using the thermodynamic cycle shown in Figure 4.7. This cycle shows that in order to calculate the relative free energy of binding between two ligands, it is necessary to perform two distinct alchemical simulations. A first simulation where a ligand L_1 is mutated into a ligand L_2 while both interact with the solvent environment (ΔG_L), and a second simulation where the ligand L_1 is mutated in the ligand L_2 in the binding site while they interact with the solvent and the protein (ΔG_{PL}). The relative binding free

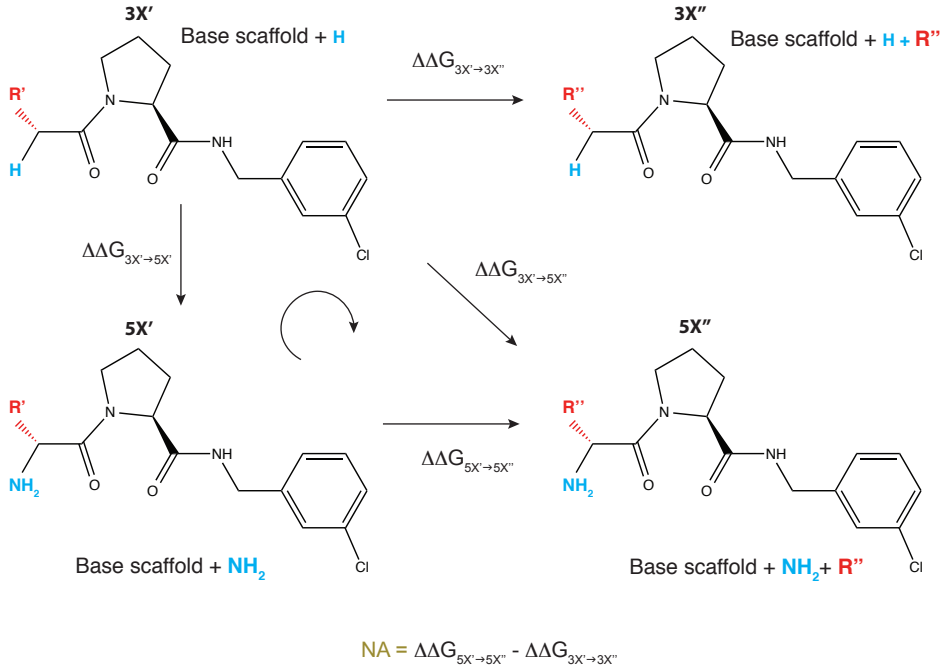


Figure 4.6: *The ligand $5X''$ presents both the modifications of the ligands $5X'$ and $3X''$. Therefore, the non-additivity can be evaluated subtracting from $\Delta\Delta G_{3X'\to 5X''}$ the sum between $\Delta\Delta G_{3X'\to 5X'}$ and $\Delta\Delta G_{3X'\to 3X''}$. Using the highlighted thermodynamics cycle the non additivity equates the difference between the relative free energy of binding $\Delta\Delta G_{5X'\to 5X''}$ and $\Delta\Delta G_{3X'\to 3X''}$*

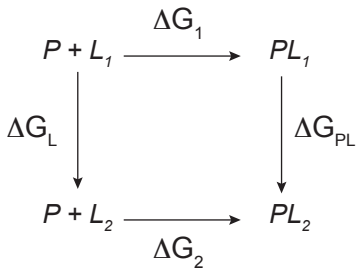


Figure 4.7: *The thermodynamic cycle used to calculate the relative free energy of binding ($\Delta\Delta G = \Delta G_2 - \Delta G_1 = \Delta G_{PL} - \Delta G_L$) between two ligands L_1 and L_2.*

energies calculated for the 3 and 5 series are reported in the relative free energy map Figure 4.8. The ligand B was selected as reference end state most of the time

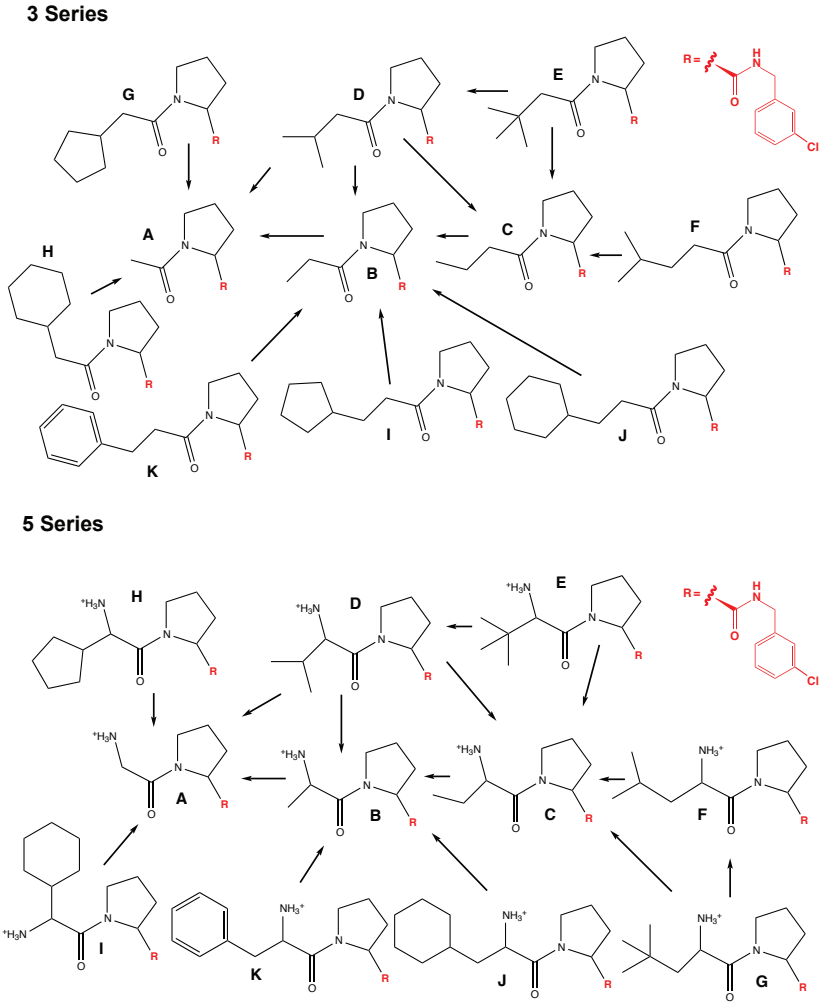
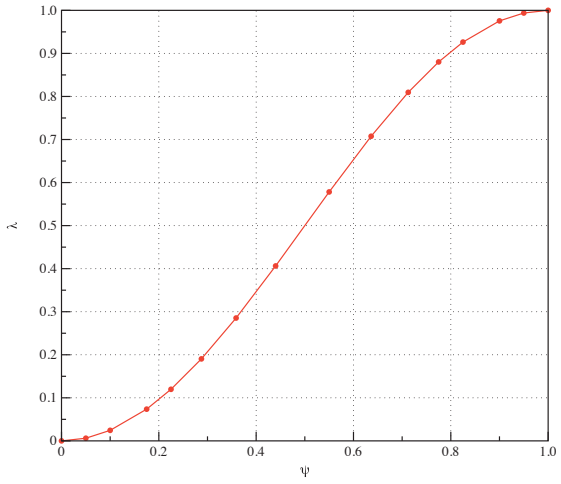


Figure 4.8: *The relative free energy map used to calculate relative free energy of binding between the ligands in the* ***3*** *(*$X = H$*) and* ***5*** *(*$X = NH_2$*) series.*

because this scaffold is central to the web of defined alchemical transformations. The free energy changes related to the complex state ΔG_{PL} and the solvent state ΔG_L (Figure 4.7) were calculated using the implemented free energy code described in the Chapter two and based on the alchemical transformation method. The coupling parameter λ used to mutate a ligand into another was modulated in the range $[0, 1]$ where 0 denotes the state where the ligand is the starting state and 1 denotes the final ligand. The transformations were performed selecting 16

λ values: (0.00000, 0.00616, 0.02447, 0.07368, 0.11980, 0.19045, 0.28534, 0.40631, 0.57822, 0.70755, 0.80955, 0.88020, 0.92632, 0.97553, 0.99384 and 1.00000). These values were generated by using the Chebyshev technique described in the second chapter by using equation 2.15. Figure 4.9 shows the generated windows for the selected value of ψ (equation 2.15) and it is interesting to observe how the point density increases nearby the boundaries of the integration region where high variations of the free energy gradient are usually observed.



ψ	λ
0.00000	0.00000
0.05000	0.00616
0.10000	0.02447
0.17500	0.07368
0.22500	0.11980
0.28750	0.19045
0.35875	0.28534
0.44000	0.40631
0.55000	0.57822
0.63625	0.70755
0.71250	0.80955
0.77500	0.88020
0.82500	0.92632
0.90000	0.97553
0.95000	0.99384
1.00000	1.00000

Figure 4.9: *The selected values of the coupling parameters λ. The values were generated by using the Chebyshev node technique (§ 2.4) with the aim of reducing the polynomial regression instability as the degree of the polynomial is increased.*

The bonded and non-bonded force field parameters involved in the mutations were linearly interpolated between the starting and final ligand parameters. In order to circumvent steric clashes at the end points of the simulations the soft core potential[134] was used between atoms that can be created or annihilated as described in Chapter two. The coulomb power and the delta shift soft-core parameters were respectively set to 0 and 2. Free energy changes were calculated by using the FDTI method and setting the delta increment to $\Delta\lambda = 10^{-3}$. The TI integral was numerically estimated by using a polynomial interpolation of seventh-order. For each window the ensemble average was calculated by sampling the system using MD. In the production run each window was sampled for 5 and 10 ns using the NPT ensemble and setting the pressure and the temperature

respectively to 1 atm and 300 K. The pressure was regulated by using Monte Carlo Barostat [101;102] with an update frequency of 25 MD steps. The Andersen Thermostat [82] was used to keep the temperature constant, selecting a collision frequency of 10 ps^{-1}. The simulations were performed by using the Leapfrog-Verlet integrator with a 2 fs time step. All the bonds were constrained to their equilibrium distance and the non-bonded interactions were evaluated by using an atom based cut off scheme setting the cutoff distance to 10 Å . The electrostatic interactions were calculated by using reaction field with the medium dielectric constant set to the water dielectric constant ($\epsilon_{solvent} = 78.3$). A total of $5 \cdot 10^4$ gradient values were collected in 10 ns simulation and each complex state calculation was repeated three times. At the beginning of each run the particle velocities were randomly generated accordingly to the Maxwell-Boltzmann distribution at 300 K temperature. The uncertainties were estimated as standard error deviation of the mean over the three independent runs. The calculation in the solvent state were performed for 5 and 10 ns one time and the uncertainties were estimated using block averaging. In order to circumvent steric clashes at beginning of the production run due to the equilibration stage performed on the starting mutant only ($\lambda = 0$), each window was re-minimized for 500 steps and re-equilibrated. This stage was performed by changing the coupling parameter between 0 and the selected window in steps of 0.1. For each one of these values an equilibration of 2 ps with 0.5 fs time step was performed setting the pressure and temperature respectively to 1 atm and 300 K. At the end of the re-equilibration stage the time step was re-set to 2 fs and no data was collected during this period. All the production runs were performed on a GPU cluster using nVidia K20 and M2090 graphic card units. During the runs the OpenMM release 5.2 was used by setting the GPU platform to OpenCL.

4.3 Free energy prediction

Predicted relative free energies of binding are reported in Table 4.1, along with the experimental data. In order to validate the correct convergence, 5 ns and 10 ns MD simulations were performed for each relative binding affinity calculation.

Table 4.1: *Experimental (Exp) and Predicted (Pred) relative free energy of binding between the ligands in the 3 and 5 series. The predictive relative free energies are shown for 5 and 10 ns MD simulation time used to sample the systems. Results are reported in kcal/mol.*

	Exp	**Pred 5ns**	**Pred 10ns**
3B → 3A	0.7 ± 0.4	0.52 ± 0.08	0.59 ± 0.02
3C → 3B	0.2 ± 0.2	0.50 ± 0.07	0.63 ± 0.04
3D → 3A	1.3 ± 0.4	1.2 ± 0.2	1.3 ± 0.1
3D → 3B	0.6 ± 0.2	0.9 ± 0.2	1.2 ± 0.1
3D → 3C	0.4 ± 0.2	0.4 ± 0.1	0.49 ± 0.06
3E → 3C	−0.24 ± 0.06	0.8 ± 0.1	0.80 ± 0.06
3E → 3D	−0.7 ± 0.2	0.8 ± 0.2	0.48 ± 0.05
3F → 3C	0.6 ± 0.2	0.7 ± 0.1	1.07 ± 0.05
3G → 3A	1.8 ± 0.4	2.4 ± 0.5	2.9 ± 0.1
3H → 3A	2.2 ± 0.4	2.9 ± 0.5	3.0 ± 0.2
3I → 3B	0.9 ± 0.2	3.5 ± 0.5	2.47 ± 0.03
3J → 3B	1.6 ± 0.2	4.4 ± 0.5	4.3 ± 0.2
3K → 3B	0.9 ± 0.2	2.5 ± 0.7	2.4 ± 0.5
5B → 5A	−0.8 ± 0.1	−0.6 ± 0.2	0.0 ± 0.2
5C → 5B	0.9 ± 0.1	1.1 ± 0.2	0.7 ± 0.02
5D → 5A	0.7 ± 0.2	1.4 ± 0.3	2.4 ± 0.1
5D → 5B	1.5 ± 0.1	1.6 ± 0.3	1.49 ± 0.08
5D → 5C	0.6 ± 0.1	0.37 ± 0.08	0.99 ± 0.03
5E → 5C	1.2 ± 0.1	1.0 ± 0.1	1.17 ± 0.03
5E → 5D	0.6 ± 0.1	0.7 ± 0.1	0.98 ± 0.03
5F → 5C	1.5 ± 0.2	0.1 ± 0.2	0.84 ± 0.09
5G → 5C	1.7 ± 0.1	0.26 ± 0.04	0.28 ± 0.03
5G → 5F	0.2 ± 0.2	0.0 ± 0.3	0.7 ± 0.3
5H → 5A	1.7 ± 0.2	3.6 ± 0.4	3.2 ± 0.2
5I → 5A	2.4 ± 0.2	5.1 ± 0.4	4.9 ± 0.2
5J → 5B	3.4 ± 0.2	4.6 ± 0.3	4.4 ± 0.4
5K → 5B	3.1 ± 0.2	3.9 ± 0.8	2.7 ± 0.3

Table 4.2 reports the relative free energy of binding selecting the ligand B as reference state in both series which will be assumed from now on unless otherwise stated and Figure 4.10 shows the comparison between experimental and predicted results. For some ligands the calculation of the relative free energy of binding

Table 4.2: *Experimental and Predicted relative free energy of binding selecting B as reference state in both series ($3B \rightarrow 3X$ and $5B \rightarrow 5X$). The predicted values are shown for 10 ns only. For some ligands the relative free energy was averaged using different thermodynamic paths and reported where needed. Data is shown in kcal/mol*

	Exp	**Pred**
3A	0.7 ± 0.4	0.59 ± 0.02
3B	0.0 ± 0.0	0.0 ± 0.0
3C	−0.2 ± 0.2	−0.63 ± 0.04
3D	−0.6 ± 0.2	−1.2 ± 0.1
3E (B→D→E ; B→C→E)	0.0 ± 0.2	−1.53 ± 0.07
3F (B→C→F)	−0.8 ± 0.2	−1.70 ± 0.07
3G (B→A→G)	−1.1 ± 0.6	−2.3 ± 0.1
3H (B→A→H)	−1.5 ± 0.6	−2.4 ± 0.2
3I	−0.9 ± 0.2	−2.47 ± 0.03
3J	−1.6 ± 0.2	−4.3 ± 0.2
3K	−0.9 ± 0.2	−2.4 ± 0.5
5A	−0.8 ± 0.1	0.0 ± 0.2
5B	0.0 ± 0.0	0.0 ± 0.0
5C	−0.9 ± 0.1	−0.76 ± 0.02
5D	−1.5 ± 0.1	−1.49 ± 0.08
5E (B→C→E ; B→D→E)	−2.1 ± 0.1	−2.20 ± 0.05
5F (B→C→F)	−2.4 ± 0.2	−1.60 ± 0.09
5G (B→C→G ; B→C→F→G)	−2.6 ± 0.2	−1.7 ± 0.2
5H (B→A→H)	−2.6 ± 0.2	−3.1 ± 0.3
5I (B→A→I)	−3.2 ± 0.3	−4.9 ± 0.3
5J	−3.4 ± 0.2	−4.4 ± 0.4
5K	−3.1 ± 0.2	−2.8 ± 0.3

was calculated averaging the relative free energy changes along different possible paths in the relative free energy map Figure 4.8. For example for the ligand $3E$ two paths were selected: $3B \rightarrow 3D \rightarrow 3E$ and $3B \rightarrow 3C \rightarrow 3E$. Table 4.2 reports the selected paths for each ligand where needed. In order to quantify the level of agreement between experimental and predicted data, three different statistical quantities were calculated: the Coefficient of Determination (R^2), the Mean Unsigned Error (MUE) and the Predictive Index (PI). The Predictive

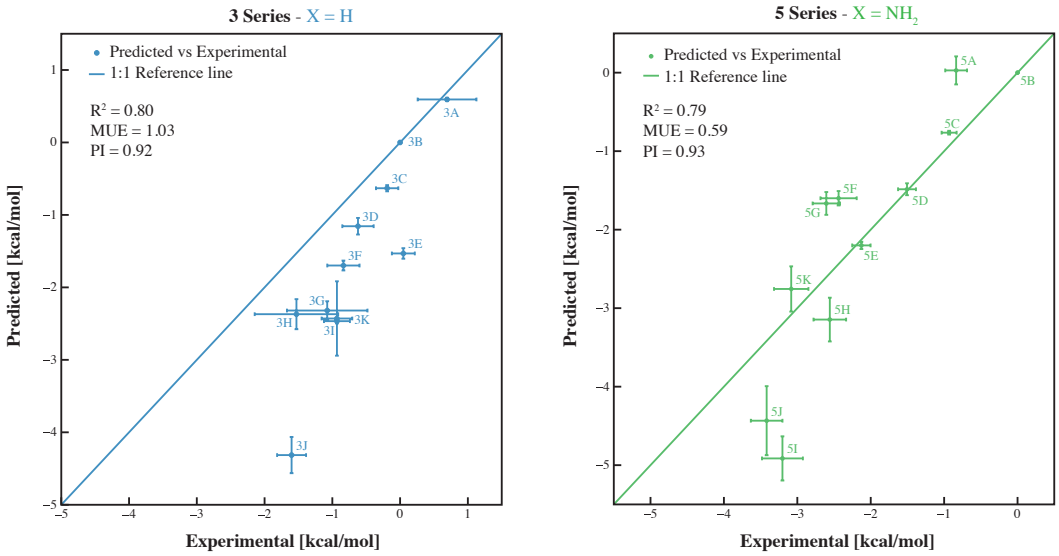


Figure 4.10: *Experimental versus Predicted relative free energy of binding selecting B as reference state in both series. The predicted values are shown for 10 ns only. Three different statistical quantities were calculated to assess the agreement between the data: R^2, MUE and PI. The continue line represents the perfect theoretical agreement.*

Index[67] is used to rank compounds in potency order. This index is defined as follows:

$$PI = \frac{\sum_{j>i} \sum_i w_{ij} C_{ij}}{\sum_{j>i} \sum_i w_{ij}} \ , \tag{4.7}$$

where

$$w_{ij} = |E(i) - E(j)| \tag{4.8}$$

and

$$C_{ij} = \begin{cases} -1 & \Longleftrightarrow \ \frac{E(j)-E(i)}{P(j)-P(i)} < 0 \\ 0 & \Longleftrightarrow \ P(j) - P(i) = 0 \\ +1 & \Longleftrightarrow \ \frac{E(j)-E(i)}{P(j)-P(i)} > 0 \end{cases} \tag{4.9}$$

In the previous equations, the term $E(i)$ is the experimental binding affinity of compound i and $P(i)$ its predicted value. In case of perfect ranking the PI index is +1 while in case of completely wrong ranking is -1 and a value of 0 represents a totally random ranking. The overall agreement between experimental and pre-

dicted data was quite satisfactory as shown in Figure 4.10 and on average the predicted data underestimates the experimental data. In addition the agreement is slightly better with the 5 series compared to the 3 series as shown by R^2, MUE and PI calculations.

In molecular simulation one of the most significant aspects to monitor is the poor system sampling which could potentially lead to wrong convergence and reproducibility issues. In order to check the result convergence, the relative binding free energy calculations were simulated in complex for 5 and 10 ns three times per each window. The result analysis showed that the the variations between 5 and 10 ns did not produce significant changes in the calculated binding free energies in most systems excepted for some more problematic transformations. In addition, the solvent simulations produced very close results extending the situation time, in some cases up the second decimal place and, therefore, the solvent simulations were simulated one time only and estimating their uncertainties using block averaging with a confidence interval of 95%. Overall, the maximum uncertainty on the relative binding affinity calculated as standard error of the mean along the triplicates was less than 0.5 kcal/mol produced in some problematic simulations. The higher discrepancy in the relative free energy of binding in the 3 series is produced by the ligand J while in the 5 series by the ligands J and I. In most of the cases in these simulations the alchemical transformation involved a relatively high number of atoms such as for the ligand 5J. In this mutation 14 atoms were simultaneously transformed from a full interactive to a non-interactive atoms producing poor convergence issues.

With regard to correctly quantify the level of agreement between experimental and predicted data an error analysis was conducted on the selected statistical quantities R^2, MUE and PI. In this analysis the experimental e_i and the predicted p_i relative binding free energies related to the ligand $i \in \{1, \dots, N\}$ (N is the total number of the ligands) can be represented as the sets $\{\dots, e_i \pm \Delta e_i, \dots\}$ and $\{\dots, p_i \pm \Delta p_i, \dots\}$, where Δe_i and Δp_i are respectively the experimental and predicted statistical uncertainties. It is often assumed that the relative binding affinity is normal distributed and, therefore, for each data point e_i and p_i it is

possible to generate new data points $\widetilde{e_i}$ and $\widetilde{p_i}$ drawing from a normal distribution $N(\mu, \sigma^2)$ i.e. $\widetilde{e_i} \sim N(\mu = e_i, \sigma^2 = \Delta e_i)$ and $\widetilde{p_i} \sim N(\mu = p_i, \sigma^2 = \Delta p_i)$. For each pair sets $\{\dots, (e_i, \widetilde{e_i}), \dots\}$ (Experimental-Experimental), $\{\dots, (p_i, \widetilde{p_i}), \dots\}$ (Predicted-Predicted) and $\{\dots, (e_i, \widetilde{p_i}), \dots\}$ (Experimental-Predicted) it is possible to determine R^2, MUE and PI and, the procedure can be re-iterated n times drawing new $\widetilde{e_i}$ and $\widetilde{p_i}$ values from a normal distribution. Therefore, R^2, MUE and PI can be represented as statistical variables themselves with probability distributions f_{R^2}, f_{MUE} and f_{PI}. The drawing procedure aims to numerically simulate both the experimental and computed binding affinity measurements for each ligand many times. This is useful to quantify the experimental and predicted result reliability by estimating, for instance, an error interval to quantify the spread of these distributions. A possible interval can be computed by using the cumulative probability F related to each one the previous distribution. The interval:

$$a \leq \bar{X} \leq b \ , \tag{4.10}$$

where $a = F^{-1}(0.03)$, $b = F^{-1}(0.98)$ and $\bar{X}$ is the average value calculated for R^2, MUE and PI was considered for this error analysis. In it, the probability to find one of the selected statistical variables is 95%, which represents nearly the entire population. The described error analysis procedure selecting $n = 10^6$ generates the results shown in Table 4.3 and Figure 4.11 The error analysis on the Experimental-Experimental data (Figure 4.3 (a)) showed that the 5 series presents improved statistics (R^2, MUE and PI) compared to the 3 series. This could indicate that the 5 series is experimentally more reliable than the 3 series. The error analysis conducted on the Predicted-Predicted data (Figure 4.3 (b)) showed that there are not significant discrepancies between the 5 and 3 series while the most significant result obtained from the error analysis conducted on the Experimental-Predicted data (Figure 4.3 (c)) showed that the MUE error is higher in the 3 series compared to the 5 series while the PI index in the 5 series is higher than the 3 series. In addition, the R^2, MUE and PI values calculated from the experimental and predicted data and shown in Figure 4.10 are in the

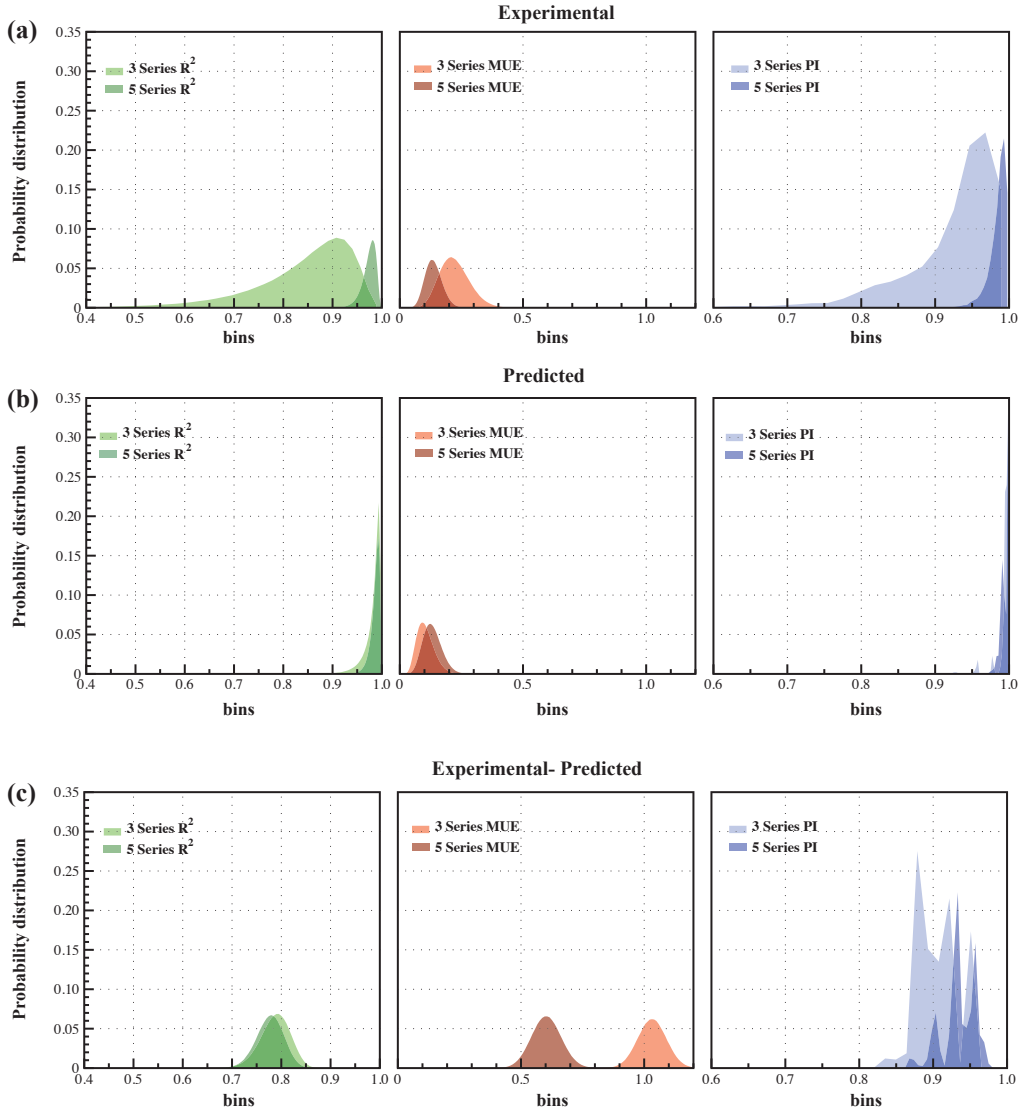


Figure 4.11: *The probability distributions of R^2, MUE and PI for the 3 and 5 series. Each distribution was simulated drawing 10^6 times from a normal distribution related to the experimental, predicted and experimental-predicted data.*

Table 4.3: *Experimental, predicted and experimental-predicted error analysis for the determination coefficient, the mean unsigned error and the predictive index. The analysis was conducted selecting an interval where the probability to find R^2 MUE or PI is 95% and drawing 10^6 times from a normal distribution.*

Experimental		
	3 Series	**5 Series**
$\mathbf{R^2}$	$0.80 \leq 0.84 \leq 0.96$	$0.94 \leq 0.97 \leq 0.99$
MUE	$0.12 \leq 0.22 \leq 0.36$	$0.11 \leq 0.14 \leq 0.21$
PI	$0.76 \leq 0.92 \leq 0.98$	$0.94 \leq 0.95 \leq 0.99$
Predicted		
	3 Series	**5 Series**
$\mathbf{R^2}$	$0.97 \leq 0.99 \leq 0.99$	$0.96 \leq 0.99 \leq 0.99$
MUE	$0.07 \leq 0.09 \leq 0.15$	$0.06 \leq 0.13 \leq 0.22$
PI	$0.97 \leq 0.99 \leq 0.99$	$0.98 \leq 0.99 \leq 0.99$
Experimental-Predicted		
	3 Series	**5 Series**
$\mathbf{R^2}$	$0.72 \leq 0.79 \leq 0.84$	$0.72 \leq 0.77 \leq 0.83$
MUE	$0.92 \leq 1.03 \leq 1.15$	$0.45 \leq 0.60 \leq 0.72$
PI	$0.85 \leq 0.90 \leq 0.95$	$0.88 \leq 0.93 \leq 0.96$

same range interval estimated by using the Experimental-Predicted distribution and reported in Table 4.3

Binding Affinity is usually predicted by using molecular docking approaches when the number of hits to validate is extremely high. These methodologies frequently involve two distinct tasks. A first stage where configurational states are generated by minimising the interaction energy between the ligand-poses and the receptor, and a second stage where the binding free energy of the complex is estimated using scoring functions. The scoring functions are generally based on an additive functional form related to intra- and inter- energetic molecular terms and the binding affinity is a weighed sum where the weighted coefficients are optimised by using different techniques such as multivariate regression analysis, genetic algorithms or artificial neural networks[135;115;136]. However, most of the time, the calculated results are frequently inaccurate compared to more robust methodologies. On the other hand, these approaches are extremely fast and therefore popular. In order to compare the accuracy of the implemented relative free energy code with a more typical scoring function software, the relative binding free energy of the selected thrombin inhibitors was also calculated by using Autodock

Vina[115]. A grid of 20 Å × 20 Å× 20 Å was centred around the thrombin binding site and a docking pose was generated for each ligand. In the docking state, the protein was rigid. In all instances where comparison was possible, it was verified that Vina produced a binding pose very similar to the available experimental data. The Vina predicted binding affinity of the best scoring pose was used for comparison.

The experimental and predict results using Vina are in Table 4.4 and in Figure 4.12 along with the determined R^2, MUE and PI. The Vina relative binding free energy predictions are quite satisfactory considering the relatively short computational time compared to the free energy calculations. Furthermore, in the 3 series the Vina MUE estimation outperformed the MUE calculated using the free energy implementation.

Table 4.4: *Experimental and Predicted relative free energy of binding selecting B as reference state in both series using Vina. The uncertainties were considered negligible in the Vina calculations. Data is shown in kcal/mol.*

	Exp	Vina
3A	0.7 ± 0.4	-0.1
3B	0.0 ± 0.0	0.0
3C	-0.2 ± 0.2	0.0
3D	-0.6 ± 0.2	-0.4
3E	0.0 ± 0.2	-0.7
3F	-0.8 ± 0.2	-0.4
3G	-1.1 ± 0.6	-1.1
3H	-1.5 ± 0.6	-1.3
3I	-0.9 ± 0.2	-1.2
3J	-1.6 ± 0.2	-1.5
3K	-0.9 ± 0.2	-1.7
5A	-0.8 ± 0.1	0.1
5B	0.0 ± 0.0	0.0
5C	-0.9 ± 0.1	0.0
5D	-1.5 ± 0.1	-0.6
5E	-2.1 ± 0.1	-0.8
5F	-2.4 ± 0.2	-0.3
5G	-2.6 ± 0.2	-0.2
5H	-2.6 ± 0.2	-1.3
5I	-3.2 ± 0.3	-1.5
5J	-3.4 ± 0.2	-1.5
5K	-3.1 ± 0.2	-1.6

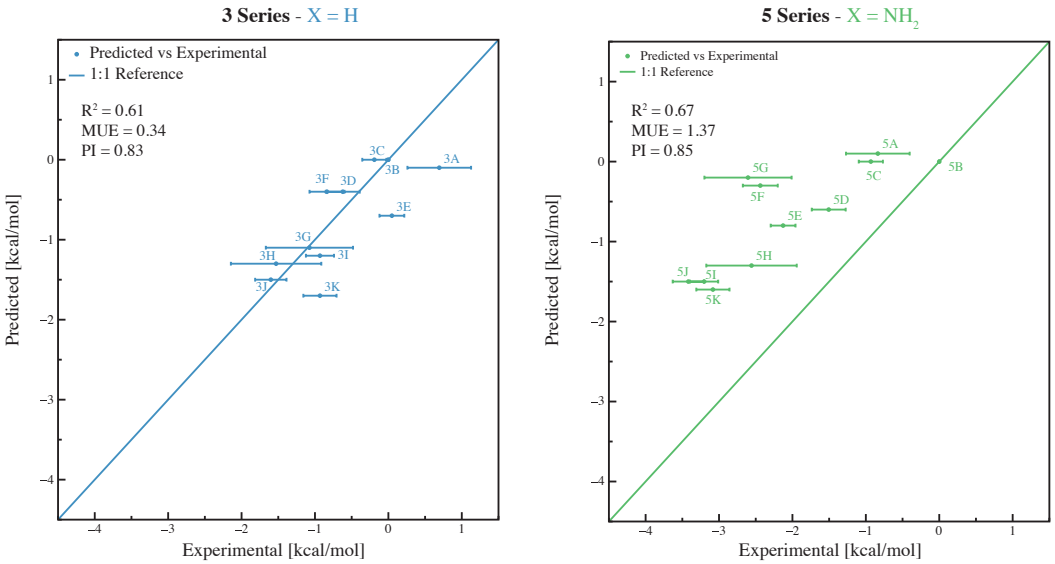


Figure 4.12: *Experimental versus Predicted relative free energy of binding selecting B as reference state in both series using Vina. The uncertainties were considered negligible in the Vina calculations.*

The non additivity levels present in the system were then finally calculated using the equation 4.6. Table 4.5 and Figure 4.13 reports the non additivity for each ligand calculated for the experimental and the predicted data using the implemented code (Sire-OpenMM) and Vina.

Table 4.5: *Experimental and Predicted non-additivity present in the system. The ligand G in the series 3 is related to the ligand H in the series 5 while the ligand H in the series 3 is related to the ligand I in the series 5 (Figure 4.8). Data is shown in kcal/mol*

	Exp	Pred	Vina
A	-1.5 ± 0.5	-0.6 ± 0.2	0.2
C	-0.7 ± 0.2	-0.1 ± 0.05	0.0
D	-0.9 ± 0.3	-0.3 ± 0.1	-0.2
E	-2.2 ± 0.2	-0.67 ± 0.09	-0.1
F	-1.6 ± 0.3	0.1 ± 0.1	0.1
G/H	-1.5 ± 0.6	-0.8 ± 0.3	-0.2
H/I	-1.7 ± 0.7	-2.5 ± 0.3	-0.2
J	-1.8 ± 0.3	-0.1 ± 0.5	0.0
K	-2.2 ± 0.3	-0.4 ± 0.4	0.1

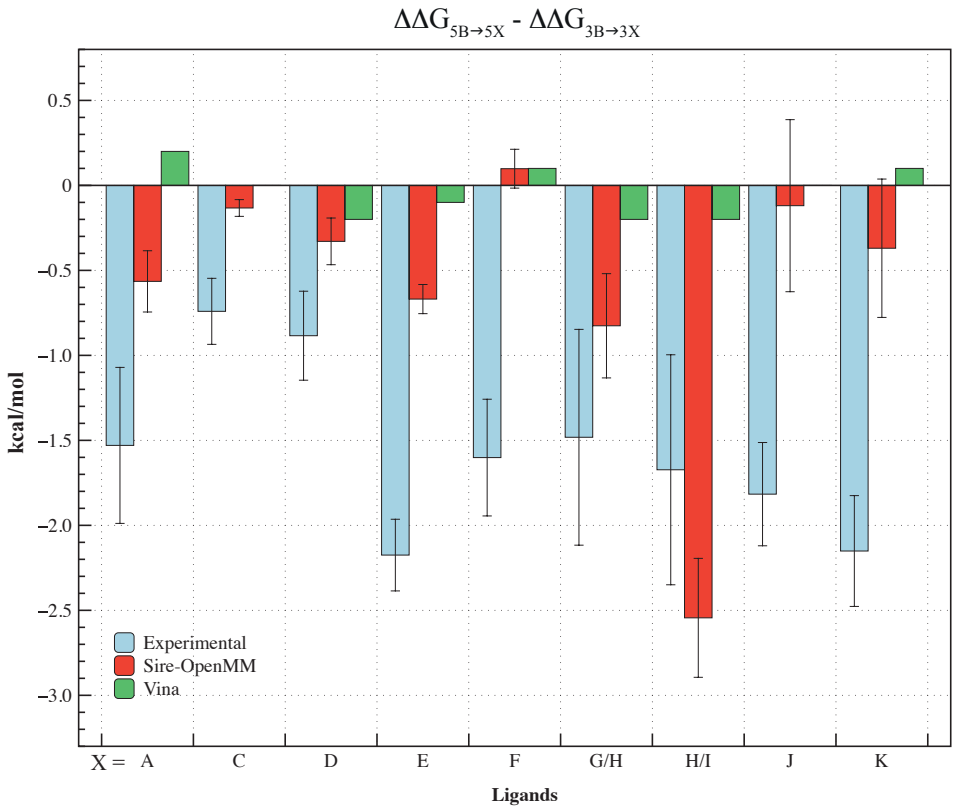


Figure 4.13: *Experimental and Predicted non-additivity levels using the implemented code Sire-OpenMM and Vina. The ligand G in the series 3 is related to the ligand H in the series 5 while the ligand H in the series 3 is related to the ligand I in the series 5.*

Figure 4.13 shows that overall Vina is not able to predict the non additivity levels for the different Thrombin inhibitors. The static model used to describe the protein by Vina prevents the capture of subtle entropic and enthalpic effects produced by the binding process and therefore docking calculations should be used with care. The predicted level of non-additivity by using the Sire-OpenMM implementation is slightly better than the Vina prediction but the overall agreement is quite poor. Probably, just in the H/I system the agreement is comparable with the experimental data. The main problem seems to be related to the 3 series. Indeed, the predicted relative binding affinities are systematically underestimated for each ligand in this series while this does not happen in the 5 series.

4.4 Free Energy Analysis

Figure 4.14 reports some significant examples of free energy gradient for the solvent and complex simulations used to compute the relative binding free energy by using the FDTI method described in Chapter two. Furthermore, the result convergence was validated by considering thermodynamic cycle closures. Indeed, theoretically along a closed thermodynamic path the free energy change must vanish and this is frequently used as a quality control parameter. In the considered mutations for the 3 and 5 series there are respectively three and four thermodynamic cycles. Figure 4.15 reports the relative binding free energy changes for each cycle. This analysis showed that the maximum discrepancy for the 3 and 5 series was respectively 0.5 kcal/mol and 1.2 kcal/mol. An interesting calculation was offered by the mutation $5B \rightarrow 5A$. Although this transformation does not involve many atoms, it seems to be a very hard alchemical mutation. Figure 4.16 reports the free energy gradient recorded along the three different runs for the simulation in complex. As shown by Figure 4.16 (a) this transformation presents a very noisy area between the windows 0 and 0.28750. With the aim of improving the convergence 12 extra windows were added in this range and simulated for 10 ns producing the result shown in Figure 4.16 (b). However, the extra windows did not improve the calculation of the relative binding affinity. In order to understand the issue, one of the window showing a significant change in the free energy gra-

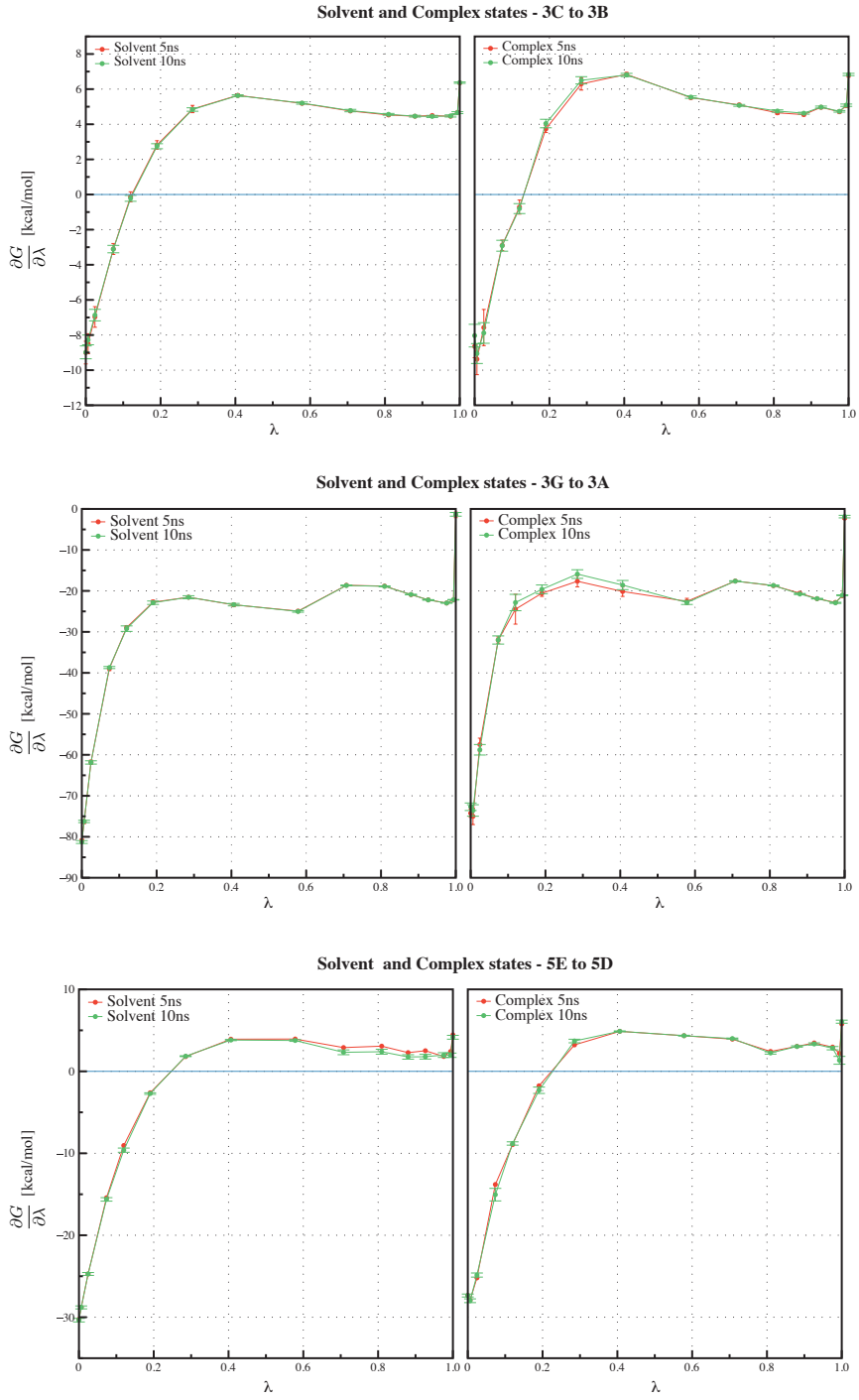


Figure 4.14: *Free energy gradients for different mutations in solvent and complex. Each selected window was simulated for 5 and 10 ns MD to monitor the gradient convergence. The free energy gradient uncertainties are plotted using block averaging selecting a confidence interval of 95%. The solvent simulations usually exhibit a very smooth and convergent behaviour. On the other hand, the simulation in complex presented poor convergence for some difficult alchemical mutations.*

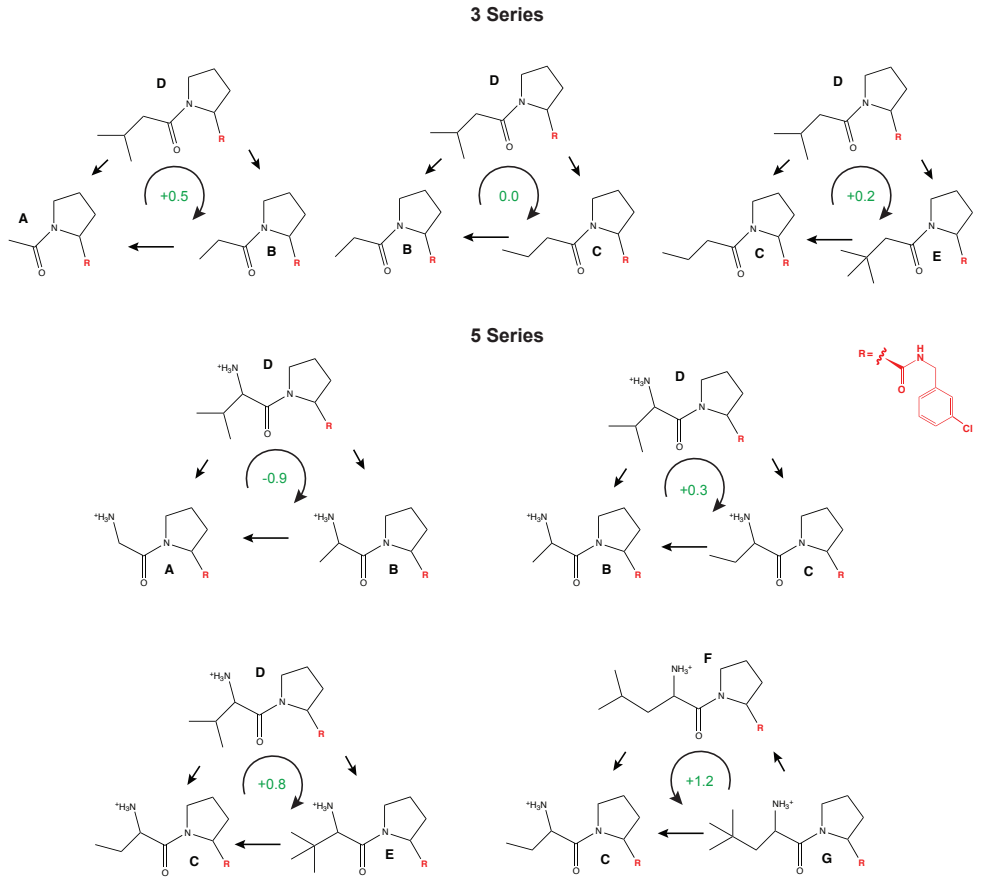


Figure 4.15: *In the selected alchemical transformations are present three thermodynamic cycles related to the 3 Series and four related to the 5 series. In the 3 series the maximum discrepancy from the perfect closure was 0.5 kcal/mol while in the 5 series 1.2 kcal/mol.*

dient between the different runs was selected for further analyses. In particular the window $\lambda = 0.19045$ presented an unsigned free energy gradient difference of nearly 5 kcal/mol between the second and third run in the simulation in complex. An analysis of the trajectories highlighted a conformational change of one dihedral angle in the mutant ligand. The starting dihedral angle between the second and third run was not the same with a difference of nearly 90 degrees and it was probably caused by the different starting random conditions along the equilibration stages for the selected window as shown in Figure 4.17 (a). At half of the third simulation the dihedral angle flipped to the same dihedral angle recorded for the

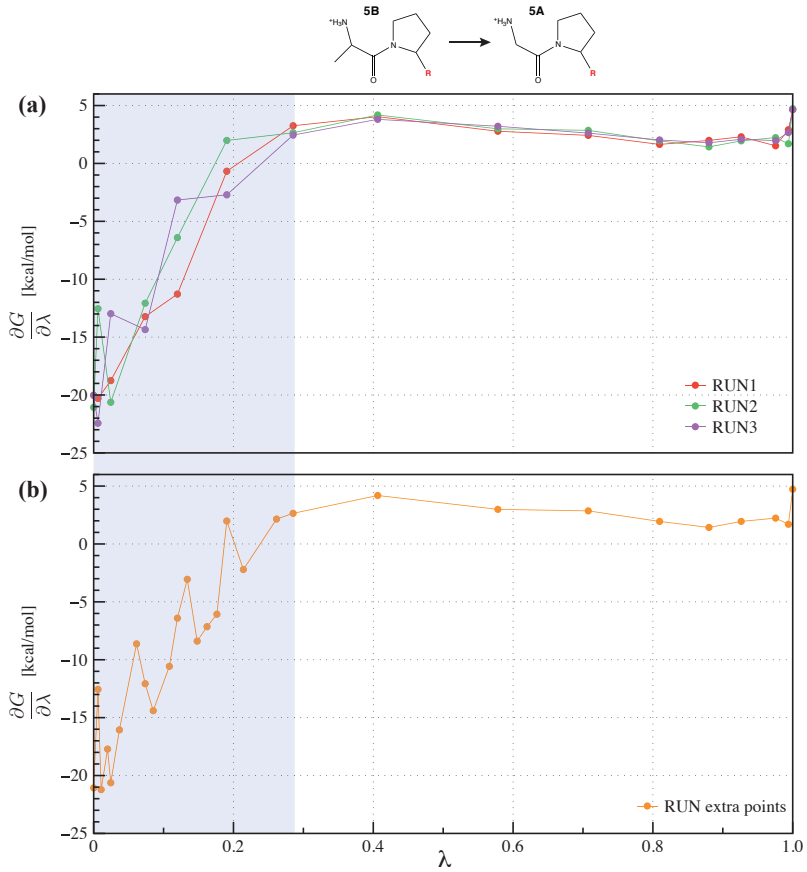


Figure 4.16: *In (a) the free energy gradient related to the mutation* $5B \rightarrow 5A$ *for the three performed simulations: RUN1, RUN2 and RUN3. In the highlighted area the window convergence was poor along the runs. In order to mitigate the issue 12 extra windows were added in this area (b) but, the gradient still presented convergence problems (RUN extra points) without any significant improvement in the overall relative free energy of binding.*

second run. The different starting conformations produced different free energy gradients as shown in Figure 4.17 (b) that eventually will converge to the same value after the dihedral angle flipping. This phenomenon was also observed for other windows involved in the noisy area.

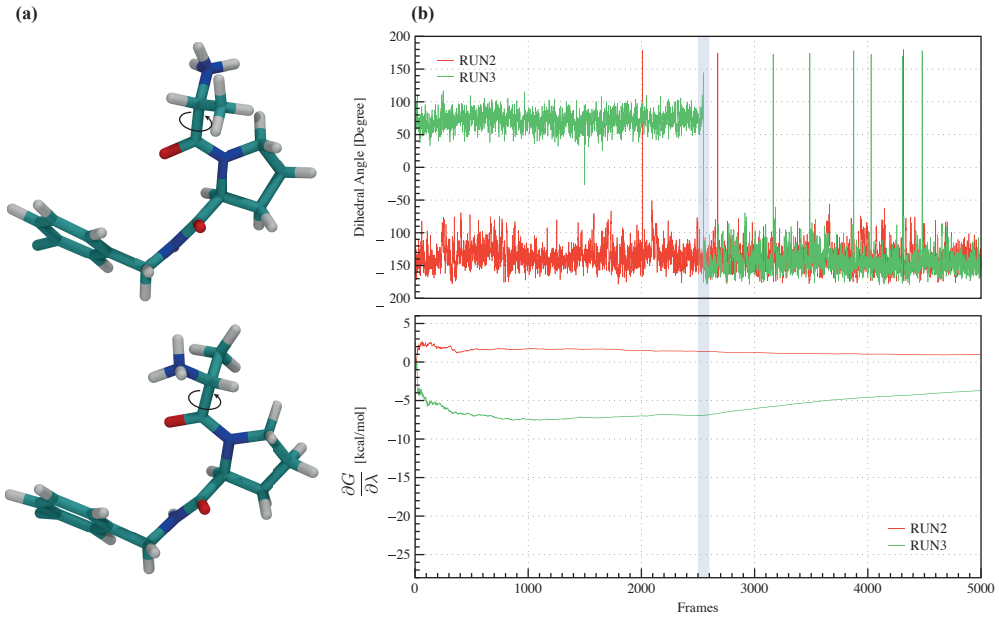


Figure 4.17: *In the mutation $5B \rightarrow 5A$ a difference of nearly 5 kcal/mol in free energy gradient was recorded for the windows $\lambda = 0.19045$ between the second (RUN2) and third simulation (RUN3) in complex. The trajectory analysis at this window showed that the issue was caused by a different starting conformation in one of the ligand dihedral angle highlighted in (a). During the third run the dihedral angle flipped to the same dihedral angle recorded in the second run. The different starting conformations between the two runs produced different free energy gradients as shown in (b) that eventually will converge to the same value after the dihedral flipping.*

It is also interesting to analyse another problematic simulation that involved the mutation $3K \rightarrow 3B$ in one of the run in complex for a particular window. The recorded free energy gradient for this run is shown in Figure 4.18 (c) for different simulation times. The window value $\lambda = 0.57822$ produced a drastic variation of the free energy gradient. The trajectory analysis emphasised the problem, which was related to one of the structural ions which was trapped by one of the

mutant carbon atoms in the benzyl ring (Figure 4.18 (a) and 4.18 (b)) producing a severe gradient change as shown in Figure 4.18 (c). A possible explanation of this phenomenon is related to the soft core parameters used and in the particular simulation dynamics. The carbon atom in the benzene ring is a "to dummy" atom and, as a consequence, the Lenard-Jones and electrostatic parameters approach to zero when the coupling parameter increase. In the selected trajectory, the VdW forces seem to be decoupled quicker than the coulomb forces. Therefore, the electrostatic attraction between the negative charged dummy carbon and the positive charged Na+ ion is not significantly opposed by the repulsive VdW forces when the carbon-ion distance become short. This particular run was repeated and the phenomenon was not observed along the other runs.

4.5 Chapter Conclusions

One of the limiting factors in the SAR stage along the critical path is the non-additivity of functional groups. In this Chapter the non-additivity of two series of congeneric inhibitors of the Thrombin protein was investigated by using molecular simulations. In particular the developed relative free energy implementation detailed in second chapter was used to predict the relative binding affinities of two investigated Thrombin inhibitor series. The predicted relative binding affinities were overall in good agreement with the experimental data as shown by the selected statistical quantities R^2 coefficient, MUE error and PI index. An error analysis was also conducted on these statistical measurements between the experimental, predicted and experimental-predicted data. The analysis showed that the experimental binding affinities of the 5 Series presented overall improved statistics compared to the 3 Series and this could indicate that this series is more reliable compared to the 3 Series. The error analysis results of the experimental-predicted data showed that the calculated statistics are in the same range of the predicted ones. Although the predicted relative binding affinities were in good agreement with the experimental data the prediction of the non-additivity levels was poor. The main problem seems to be related to the predicted relative binding affinities of the 3 Series, which systematically underestimate the experimental values. The

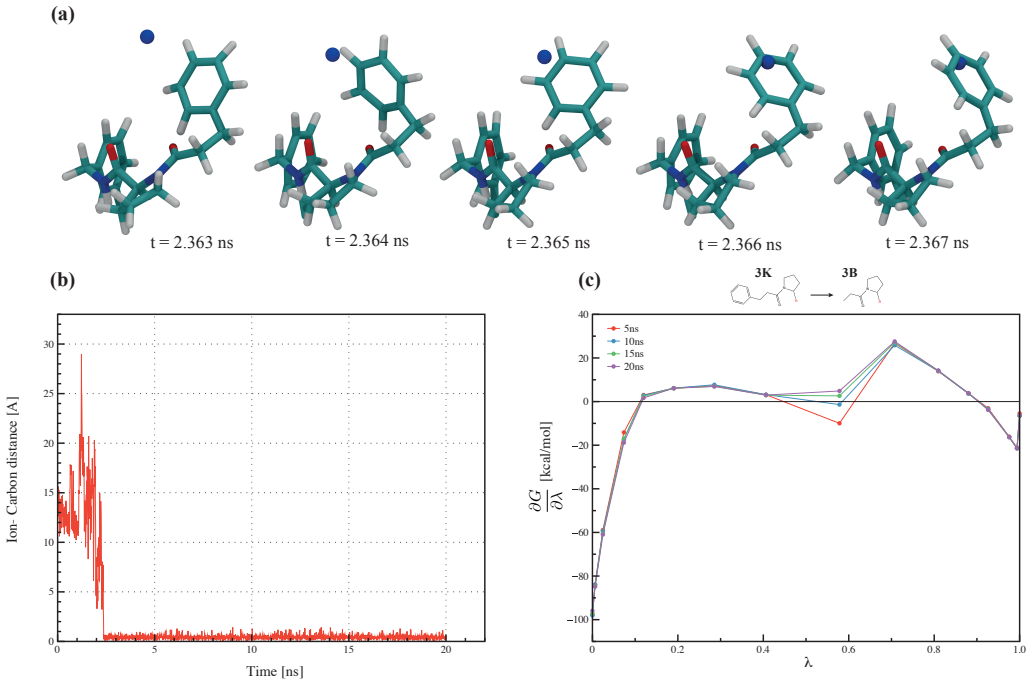


Figure 4.18: *The mutation $3K \rightarrow 3B$ presented a very interesting behaviour for the window $\lambda = 0.57822$ in one simulation. In (a) one of the structural ion Na+ was trapped by one of the mutant carbon atom in the benzyl ring along the trajectory. The event in (b) was recorded at about 2.3 ns by checking the distance between the ion and the carbon. In (c) the overall gradient presents a jump at the selected window, which is disappearing increasing the simulation time due to the accumulation of gradient values. This simulation was not considered in the analysis stage. This event is probably caused by the use of the soft core potential.*

issue could be related to force field parameterisations in this series and further investigations are required. The non-additivity levels were also predicted by using Vina, a scoring function docking program but, the level of agreement was extremely poor in this case and confirming that subtle enthalpic and entropic effect cannot be easily predicted by scoring function approaches.

"But don't you see that the whole trouble lies here? In words, words. Each one of us has within him a whole world of things, each man of us his own special world. And how can we ever come to an understanding if I put in the words I utter the sense and value of things as I see them; while you who listen to me must inevitably translate them according to the conception of things each one of you has within himself. We think we understand each other, but we never really do"

— Luigi Pirandello. Six characters in search of an author

5 Possible origins of Non-Additivity

5.1 Non-Additivity hypotheses

The correct prediction of the relative free energy of binding between the inhibitors of Thrombin is the starting point to rationalise the non-additivity origins in the system. Although, the agreement between experimental and predicted data is satisfactory as shown in Figure 4.10, the predicted non-additivity levels for the different ligands were not in systematically good agreement with the experimental data and, therefore, it was not possible to conduct an orderly analysis on each system to determine non-additivity causes. However, the predicted non additivity in the ligand pairs H/I (Figures 4.13) justifies a detailed analysis of this system only. It is worth to remember that the ligand 5I in the 5 Series corresponds to the ligand 3H in the 3 Series (apart from the extra amino group) and the predicted non-additivity level was calculated considering

the ligand B as reference state in both series as follows:

$$NA_{H/I} = \Delta\Delta G_{5B\to 5I} - \Delta\Delta G_{3B\to 3H} \tag{5.1}$$

In the previous equation, each relative binding affinity was computed along two thermodynamic paths as shown in Figure 5.1. Because of time constraints, the present analysis focused on the transformations $H/I \to A$ i.e. $5I \to 5A$ and $3H \to 3A$ only.

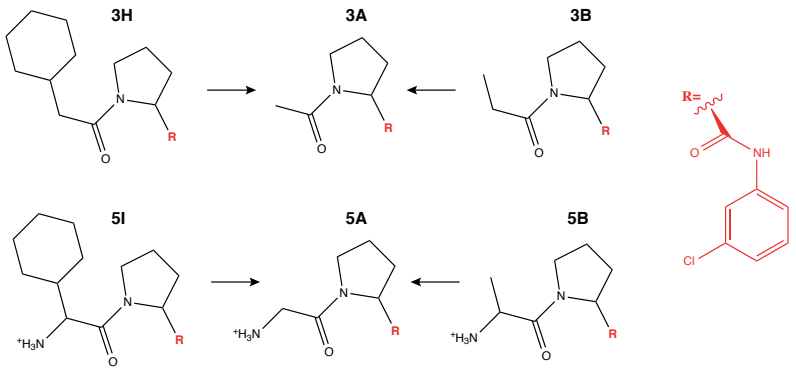


Figure 5.1: *The non-additivity in the system H/I is evaluated computing the relative binding affinities along two paths for each series. The path $5B \to 5I = 5B \to 5A \to 5I$ and $3B \to 3H = 3B \to 3A \to 3H$.*

There are many factors that may trigger the non-additivity in the Thrombin system and in particular entropic or enthalpic contributions were investigated here. An entropic effect could be generated by the additional amino group in the 5 series that may produce a protein or ligand rigidification (or both) compared to the 3 series where the amino group is missing. In addition, the positive charge of the amino group could have an enthalpic contribution related to a significant change in the electrostatic interactions between the 5 series inhibitors and the solvent compared to the 3 series. All the previous effects may cooperate together or one could be more significant than the others.

Nowadays, the main role of simulations is to predict experimental data. However, if computational models are relatively correct, it is possible to validate hypotheses that would not be possible to experimentally prove or disprove, for in-

stance, due to experimental limitation techniques. With the aim of discovering the origin of non-additivity in the selected system, it is possible to design computational experiments where the rigidity of the molecules, or the strength of ligand-solvent interactions, are artificially modified. The idea relies on the following working hypothesis:

> "if an opportune change of the ligand/protein rigidity or both or, the ligand-solvent interactions, removes the non-additivity in the system, then, the non-additivity source can be ascribed to that change"

In order to prove or disprove this hypothesis, the following computational experiments were attempted:

- *effects of protein flexibility change on the non-additivity.* If the non-additivity origin is due to a change in the protein flexibility between the ligands in the 3 and 5 series then, a protein rigidification prior binding should suppress the non-additivity;
- *effects of the ligand flexibility change on the non-additivity.* If the addition of the amino group changes the ligand flexibility then restraining all the ligands in their binding mode should suppress the non-additivity;
- *effect of solvent interactions.* If the non-additivity is caused by a change in the ligand-solvent interactions then a simulation where these interactions are limited should suppress the non-additivity.

With the aim of testing the previous hypotheses, the relative binding free energy implementation described in Chapter two was modified and extended. The first hypothesis related to the the protein rigidification can be tested by adding an artificial protein rigidification restraint. Indeed, if the amino group would change the protein rigidity it will not be able to further rigidify an already rigid protein and, therefore, the non-additivity should be suppressed. In order to rigidify the protein, a positional restraint potential was implemented and added to the relative binding free energy code. The potential was applied on selected protein atoms

and its implemented functional form is:

$$U(\mathbf{r}_i) = k_p(\mathbf{r_i} - \mathbf{r_{i0}}) \cdot (\mathbf{r_i} - \mathbf{r_{i0}}) \ , \tag{5.2}$$

where $\mathbf{r_i}$ is the position of a protein atom i, $\mathbf{r_{i0}}$ the selected space position to restraint the atom i, the constant k_p a parameter used to control the positional restraint strength and the symbol " $\cdot$ " denotes the vectorial scalar product.

In a similar way, to test the second hypothesis related to the change in ligand rigidity, restraint distances were applied between specific atom pairs. Indeed, the addition of the amino group will not be able to further rigidify an already rigid ligand and therefore the non-additivity should be suppressed if this was the cause. Atom pairs were judiciously selected between the ligand and the protein and the following restraining potential was applied between these pairs:

$$U(r_{ij}) = k_d \max(0, (r_{ij} - r_{ij}^{eq})^2 - D_{ij}^2) \ , \tag{5.3}$$

where r_{ij}^{eq} is the selected restraint distance between atom pair i and j, r_{ij} the atom pair distance, k_d a parameter used to control the restraint distance strength and D_{ij} a distance parameter used to control when apply the restraint. This approach was preferred to positional restraints to rigidify the ligands, because it allows small fluctuations between interatomic distances during the simulations.

Finally, in order to test if the non-additivity is produced by changes in the non-bonded interactions between the amino group in the 5 series and the solvent, a new potential was implemented and applied between a selected ligand atom and all the solvent atoms. The aim of the designed potential was to keep the solvent molecules far from the ligand and, therefore, to reducing the electrostatic and Lennard Jones interactions with the amino group. The functional form of the implemented potential is as follows:

$$U(r_{pj}) = k_b(r_{pj} - D_b)^2\theta(D_b - r_{pj}) \ , \tag{5.4}$$

where, p and j are respectively the selected ligand atom and a solvent atom, r_{pj}

the distance between atom p and atom j, k_b a parameter used to control the force strength applied between atom i and j, D_b a penalised solvent atom distance and θ the Heaviside function. In brief, this potential defines a spherical shield volume centred on atom p with radius D_b where solvent molecules are discouraged to enter.

5.2 Validating the non-additivity hypotheses

In the considered H/I system the non-additivity is quantified by using equation 5.1. Additionally, the thermodynamic paths $5B \rightarrow 5I$ and $3B \rightarrow 3H$ (Figure 5.1) present two single mutations each: $5B \rightarrow 5I = 5B \rightarrow 5A \rightarrow 5I$ and $3B \rightarrow 3H = 3B \rightarrow 3A \rightarrow 3H$. As a consequence, equation 5.1 can be rewritten in terms of the non-additivity along each single path as follows:

$$\begin{aligned} NA_{H/I} &= \Delta\Delta G_{5B\rightarrow 5I} - \Delta\Delta G_{3B\rightarrow 3H} = \\ &= \Delta\Delta G_{5B\rightarrow 5A} + \Delta\Delta G_{5A\rightarrow 5I} - \Delta\Delta G_{3B\rightarrow 3A} - \Delta\Delta G_{3A\rightarrow 3H} = \\ &= \Delta\Delta G_{5B\rightarrow 5A} - \Delta\Delta G_{3B\rightarrow 3A} + \Delta\Delta G_{5A\rightarrow 5I} - \Delta\Delta G_{3A\rightarrow 3H} = \\ &= NA_1 + NA_2 \ , \end{aligned} \tag{5.5}$$

where $NA_1 = \Delta\Delta G_{5B\rightarrow 5A} - \Delta\Delta G_{3B\rightarrow 3A}$ and $NA_2 = \Delta\Delta G_{5A\rightarrow 5I} - \Delta\Delta G_{3A\rightarrow 3H}$ are the single non-additivity contributions along each path. The hypotheses related to the rigidity and/or ligand-solvent interactions changes have as final goal to suppress the non-additivity in the system. Therefore, the previous equation must suppress $NA_{H/I} = 0$ which, in terms of single non-additivity components NA_1 and NA_2 is translated into the following statement:

$$NA_{H/I} = 0 \iff NA_1 = 0 \ and \ NA_2 = 0 \tag{5.6}$$

The analysis in this chapter was restricted to annihilate NA_2 only. It is worth to remember that to compute NA_2 it is necessary to perform two alchemical transformations and calculate the free energy changes required to mutate the starting ligand into the final one respectively in the protein-binding site and in

solvent Figure 4.7. In order to calculate the two relative free energy changes, each system was assembled and set as detailed in Chapter § 4.2. The relative free energy calculations were performed by using the implemented code, adding the potentials detailed in the equations 5.2, 5.3 and 5.4. The main parameters used in the calculations are detailed below. The transformations were performed selecting 16 λ windows: (0.00000, 0.00616, 0.02447, 0.07368, 0.11980, 0.19045, 0.28534, 0.40631, 0.57822, 0.70755, 0.80955, 0.88020, 0.92632, 0.97553, 0.99384 and 1.00000). In order to circumvent steric clashes, a soft core potential was used between atoms that can be created or annihilated as described in Chapter two. The coulomb power and the delta shift soft-core parameters were respectively set to 0 and 2. Free energy changes were calculated by using the FDTI method setting the delta increment to $\Delta\lambda = 10^{-3}$. The TI integral was numerically estimated by using a polynomial interpolation of seventh-order. In the production run each window was sampled for 5 ns in the complex and solvent state by using the NPT ensemble setting the pressure and the temperature respectively to 1 atm and 300 K. The pressure was regulated by using the Monte Carlo Barostat [101;102] with an update frequency of 25 MD steps. The Andersen Thermostat [82] was used to keep the temperature constant selecting a collision coefficient of 10 ps^{-1}. The simulations were performed by using the Leapfrog-Verlet integrator scheme with 2 fs time step. All the bonds were constrained to their equilibrium distance and the non-bonded interactions were evaluated by using an atom based cut off scheme setting the cutoff distance to 10 Å . The electrostatic interactions were calculated by using reaction field with the medium dielectric constant set to the water dielectric constant ($\epsilon_{solvent} = 78.3$). $2.5 \cdot 10^4$ gradient values were collected in a 5 ns simulation. In order to circumvent steric clashes at beginning of the production run due to the equilibration stage performed on the starting mutant only ($\lambda = 0$), each window was re-minimized for 500 steps and re-equilibrated. This stage was performed by changing the coupling parameter between 0 and the selected window in steps of 0.1. For each one of these values an equilibration of 2 ps with a 0.5 fs time step was performed, setting the pressure and temperature respectively to 1 atm and 300 K. At the end of the re-equilibration stage the time

step was re-set to 2 fs and no data was collected. All the simulations in complex and solvent were repeated three times and the uncertainties were estimated as standard error of the mean over the three independent runs. In this chapter, the above detailed relative free energy protocol will be used to test the different non-additivity origin hypotheses and it will be referred as the relative free energy protocol.

In order to test the hypothesis related to the protein rigidification, the potential energy expression detailed in equation 5.2 was applied on selected protein atoms. In particular all the protein heavy atoms i.e. atoms with atomic mass $\geqslant 1.10$ amu were restrained to their starting position by using a force constant $k_p = 10$ kcal mol^{-1} Å^{-2} (equation 5.2). In this case the ligand free simulations were not influenced by the application of the positional restraints. A comparison between the non-additivity computed by using or not protein positional restraints is reported in Figure 5.2 while, Figure 5.3 (a) and 5.3 (b) shows the Thrombin protein by colouring its atoms by using calculated B factors respectively for the flexible and the restrained protein.

Although the experimental non-additivity level is not in agreement with the predicted data in the selected system, the addition of positional restraints on the protein heavy atoms seem able to suppress the non-additivity in the system. This partial result could corroborate the experimental considerations of Baum et al.[76] that suggested a binding site rigidification effect through B-factor analysis. However, to prove or disprove this hypothesis, the calculation should be extended to all the selected Thrombin inhibitors when their non-additivity prediction levels will be in better agreement with the experimental data.

With the aim of validating also a possible ligand rigidification effect, restraint distances implemented by using equation 5.3 were applied to selected ligand atoms (intra-ligand restraints) and between ligand and protein atoms (inter-ligand restraints) to prevent orientational changes of the ligand respect to the protein. The selected atom pairs are reported in Figure 5.4. The equilibrium distances r_{ij}^{eq} in equation 5.3 were set as the starting distances between the selected atom pairs. The restrained force constants were set to $k_d = 25$ kcal mol^{-1} Å^{-2} while D_{ij}

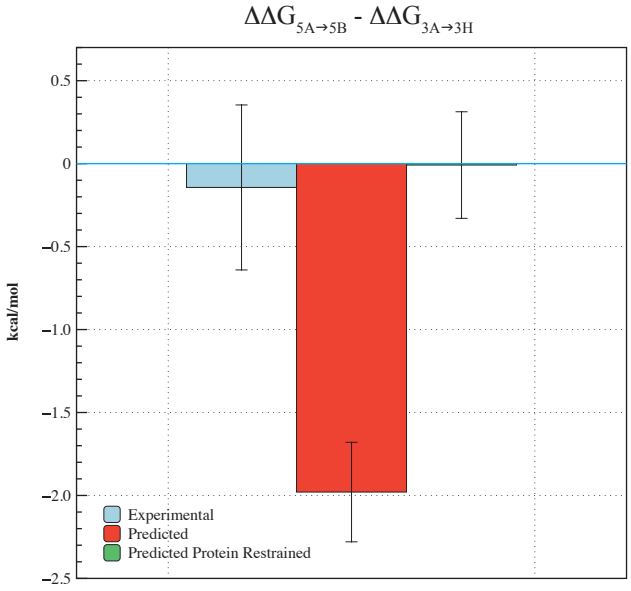


Figure 5.2: *A comparison between the Experimental (light blue), the Predicted (red) and the Predicted (green) by using positional Restraints on the protein heavy atoms in the system H/I. In both series the ligand A was selected as reference state. The application of the positional restraints seems to suppress the non-additivity in the system.*

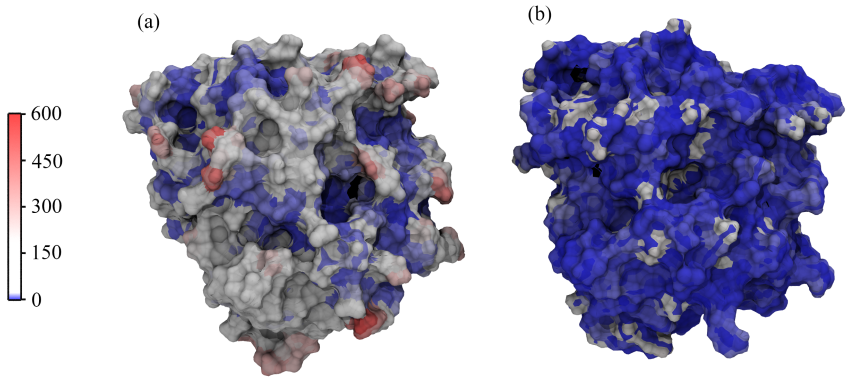


Figure 5.3: *A comparison between the B factors calculated for the un-restrained 5I complex system. (a) The B factor of the protein in complex with the ligand 5I was calculated by measuring the rmsd of each atom along the trajectory selecting as reference frame the average atomic positions recorded along the simuation. (b) The B factor calculated as in (a) but restraining the protein heavy atoms. In the scale the blue color indicates low B factor values and, therefore, rigid protein parts. The B factor was measured in* $\AA^2$ *units.*

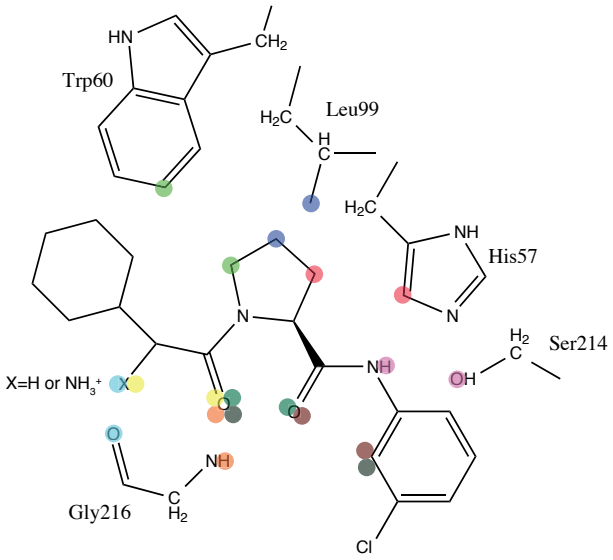


Figure 5.4: *Atom pairs selected to apply restrain distances. A restraint distance was applied between circled atoms with the same colour. Intra-ligand and inter-ligand restraint distances were selected.*

(equation 5.3) was set to 0.2 Å. The calculation of non-additivity requires also the simulations in solvent in this case and, the intra-ligand restraint distances illustrated in Figure 5.4 were applied to the ligands only. Furthermore, the protein heavy atoms in the complex simulations were restrained to their starting positions by using the positional restraint potential detailed in the equation 5.2 with the force constant $k_p = 10$ kcal mol^{-1} Å^{-2}. All the relative free energy calculations with the addition of restraint distances and positional restraint potentials were performed by using the relative free energy protocol previously detailed and, results are reported in Figure 5.5. The addition of the restraint distances seem to

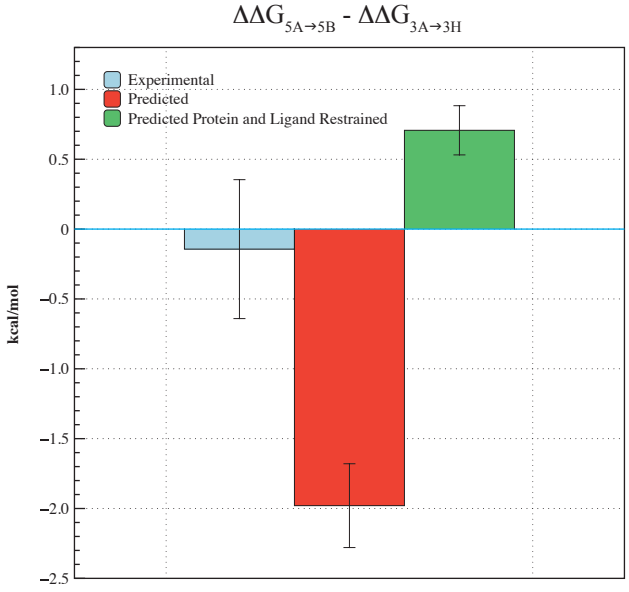


Figure 5.5: *A comparison between the Experimental (light blue), Predicted (red) and Predicted (green) by using positional Restraints on the protein heavy atoms and applying restraint distances between selected atom pairs (Figure 5.4) in the H/I system. In both series the ligand A was selected as reference state. In this case a positive cooperativity appears in the system.*

have an impact on the non-additivity, increasing its value from nearly 0.0 to +0.7 kcal/mol compared to the case with just positional restraints (Figure 5.2). However, this effect could be caused by the problematic convergence recorded for the transformation $5I \rightarrow 5A$ in the complex state as illustrated in Figure 5.6. Indeed, the window values $\lambda = 0.19045$ and $\lambda = 0.28534$ present convergence problems.

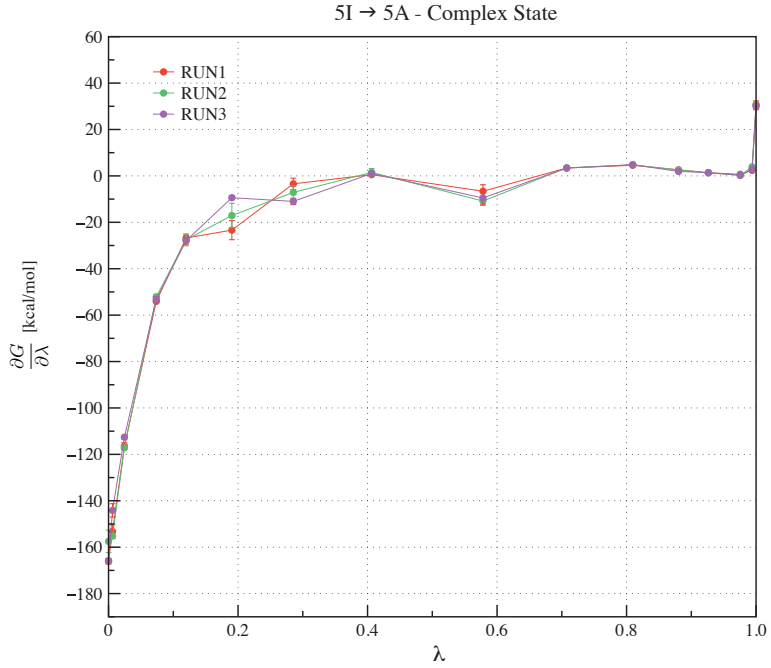


Figure 5.6: *Free energy gradient related to the* $5I \rightarrow 5A$ *alchemical mutation in complex state recorded over three independent runs. The windows* $\lambda = 0.19045$ *and* $\lambda = 0.28534$ *present poor convergence that could produce a significant error in the numerical integration and therefore a poor non-additivity estimation.*

In this case, the whole free energy gradient undergoes a significant energy change between its extreme values ($\lambda = 0$ and $\lambda = 1.0$) and poor convergence in some window values could produce a significant error in the numerical integration used to compute the free energy changes. Therefore, the non-additivity increment could be a false positive effect and further investigations are required to clarify this result.

The final hypothesis tested was related to the change in the non-bonded interactions between the amino group introduced in the 5 series and the solvent. In order to limit these interactions and suppress the non-additivity in the system, the potential in equation 5.4 was implemented. The main effect of this potential is to create a spherical volume around the ligand where solvent molecules are discouraged to enter. The simulated spherical potential was centred on the ligand atom illustrated in Figure 5.7 with a radius of $D_b = 7$ Å. The potential

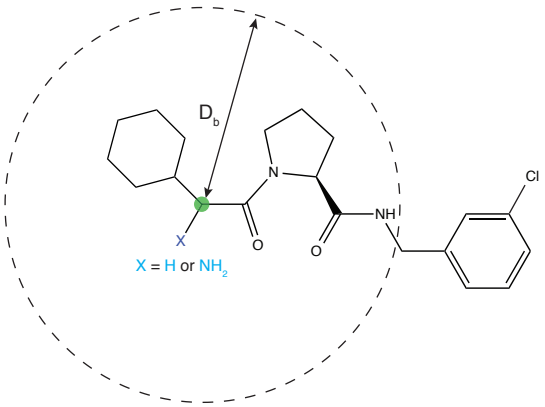


Figure 5.7: *In green the ligand atom selected to apply the potential described in equation 5.4. This atom interacts with all the solvent atoms creating a spherical volume where solvent atoms are penalised to enter and, therefore, limiting the non-bonded interactions between ligand and solvent.*

was applied on both complex and solvent simulations for the mutations $5I \rightarrow 5A$ and $3H \rightarrow 3A$. Furthermore, positional restraints and restraint distance potentials were also applied and set as detailed for the other hypotheses. The relative free energy simulations were performed by using the relative free energy protocol already described and results are reported in Figure 5.8. In order to check if the

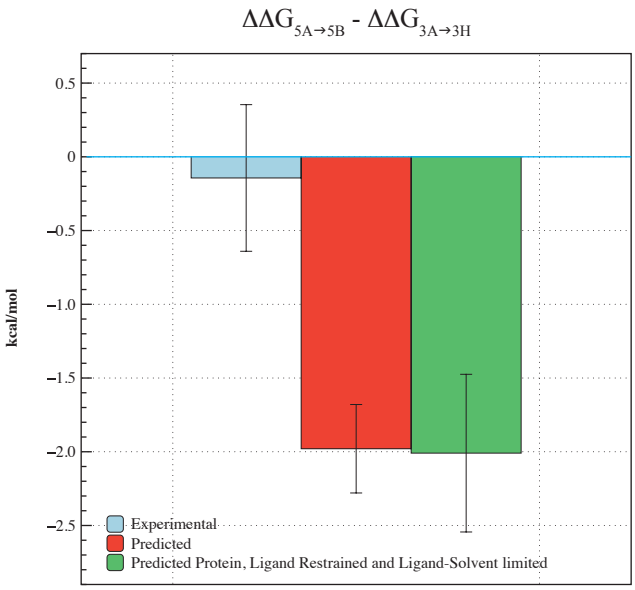


Figure 5.8: *A comparison between the Experimental (light blue), Predicted (red) and Predicted (green) by using positional Restraints on the protein heavy atoms, applying restraint distances between selected atom pairs (Figure 1.4) and creating a penalised solvent volume around the ligand in the H/I system. In both series the ligand A was selected as reference state. In this case the non-additivity seems to reappear in the system.*

spherical potential was effectively limiting the non-bonded ligand interactions the cumulative averages of the number of solvent molecules around the ligand were calculated by post-processing the simulated trajectories. Figure 5.9 reports an example of the calculated cumulative averages and Figure 5.10 reports a comparison between a system simulated with and without the application of potential energy expression 5.4.

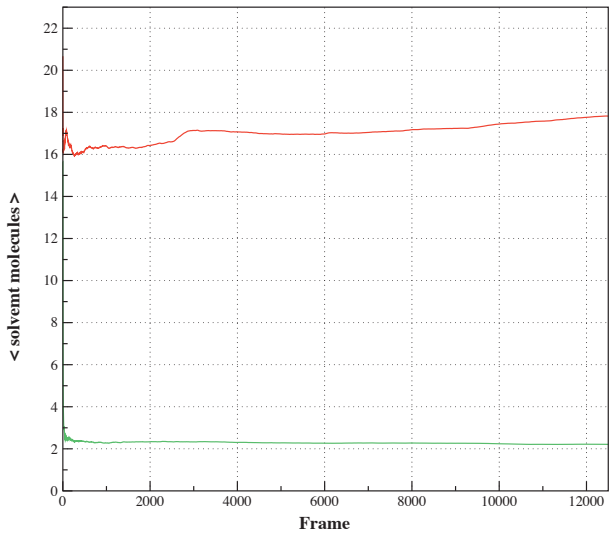


Figure 5.9: *A comparison between the cumulative average of the solvent molecule number around the ligand recorded along the complex state mutation* $5I \rightarrow 5A$ *selecting the window* $\lambda = 0.0$. *The green line represents the system where all the potentials detailed by equations 5.2, 5.3 and 5.4 are considered and the red line the system without them. The radius parameter in equation 5.4, which defines the penalised solvent spherical volume was set to* $D_b = 7$ *Å in this case.*

The predicted non-additivity level (Figure 5.8) is significant in this case and seems to reach the same level produced without the application of the potentials 5.2, 5.3 and 5.4. This behaviour can only be explained in terms of noisy and not converged alchemical simulations, which prevent the correct quantification of the non-additivity in the system. Figure 5.11 reports the different gradients recorded for all the mutations involved in the non-additivity calculations related to the system H/I considering A as reference state. In at least two mutations $3H \rightarrow 3A$ in the solvent state and, $5I \rightarrow 5A$ in the complex state there are

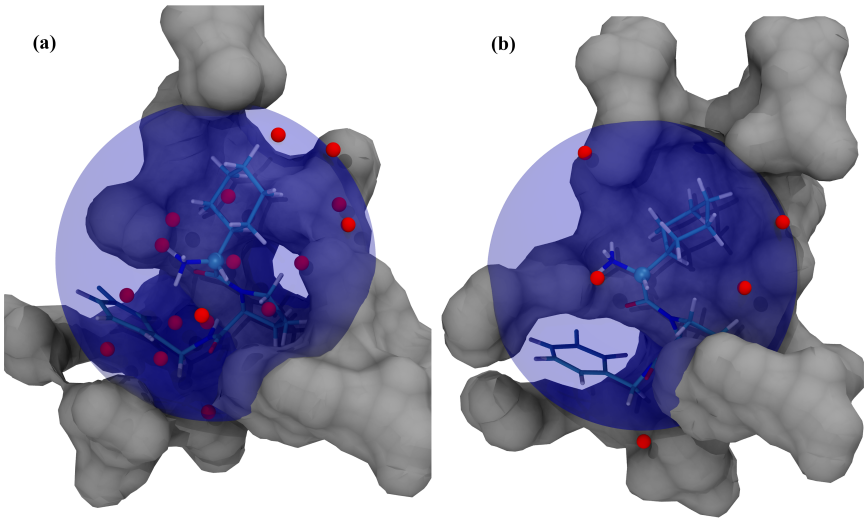


Figure 5.10: *Two snapshots extracted from two alchemical mutations of* $5I \rightarrow 5A$ *selecting the window* $\lambda = 0.0$. *In (a) and (b) the potentials 5.2, 5.3 and 5.4 where respectively not used and used during the simulations. Part of the Thrombin protein binding site is illustrated by using a grey shaded surface, while red spheres are used for oxygen water atoms. The transparent blue sphere represents the penalised solvent region. In (a) more water molecules are recorded compared to (b) as expected.*

serious convergence problems in many window values. A first recognised problem in the noisy complex simulation $5I \rightarrow 5A$ seems to be caused by the dehydration volume created around the ligand. In this volume, the positive charged amino group does not have to electrostatically interact with the polar water molecules and, in many simulations seems to form an intra-electrostatic interaction with one of the two-ligand oxygen as illustrated in Figure 5.12. Probably, selecting extra restraint distances could have mitigated this behaviour. Another source of noise in the system could be caused by the effect of the reaction forces on the ligand produced by the pressure of the water molecules. Indeed these forces are applied on the selected ligand atom to centre the sphere potential as illustrated in Figure 5.7. It is difficult to quantify this effect but, in the complex state simulations the water pressure should further buried the ligand in the binding site due to the system anisotropicity while, the effect should be less pronounced in the solvent state simulations due to the system isotropicity.

5.3 Chapter Conclusions

In this chapter a rationalisation of the non-additivity origins in the Thrombin system was attempted by analysing the non-additivity in specific inhibitors of Thrombin. Three hypotheses were tested related to an entropic effect of rigidification of the protein/ligand and an enthalpic contribution caused by changes in ligand-solvent interactions. With the aim of validating these hypotheses, three computational experiments were attempted to suppress the non-additivity in the considered systems. However, the obtained results were not able to clearly prove or disprove the hypotheses although, the protein rigidification was able to remove the non-additivity in the system. The other two hypotheses presented computational problems with the convergence of the free energy gradient and with unexpected binding mode changes to fix and future work is required to validate their reliable application to elucidate non-additivity. This chapter is a first effort to rationalise the origins of non-additivity in Thrombin, while non-additivity was suppressed with the use of a rigid protein protocol, more work is needed to confirm this effect also applies to every ligand in the considered data set. The following chapter will

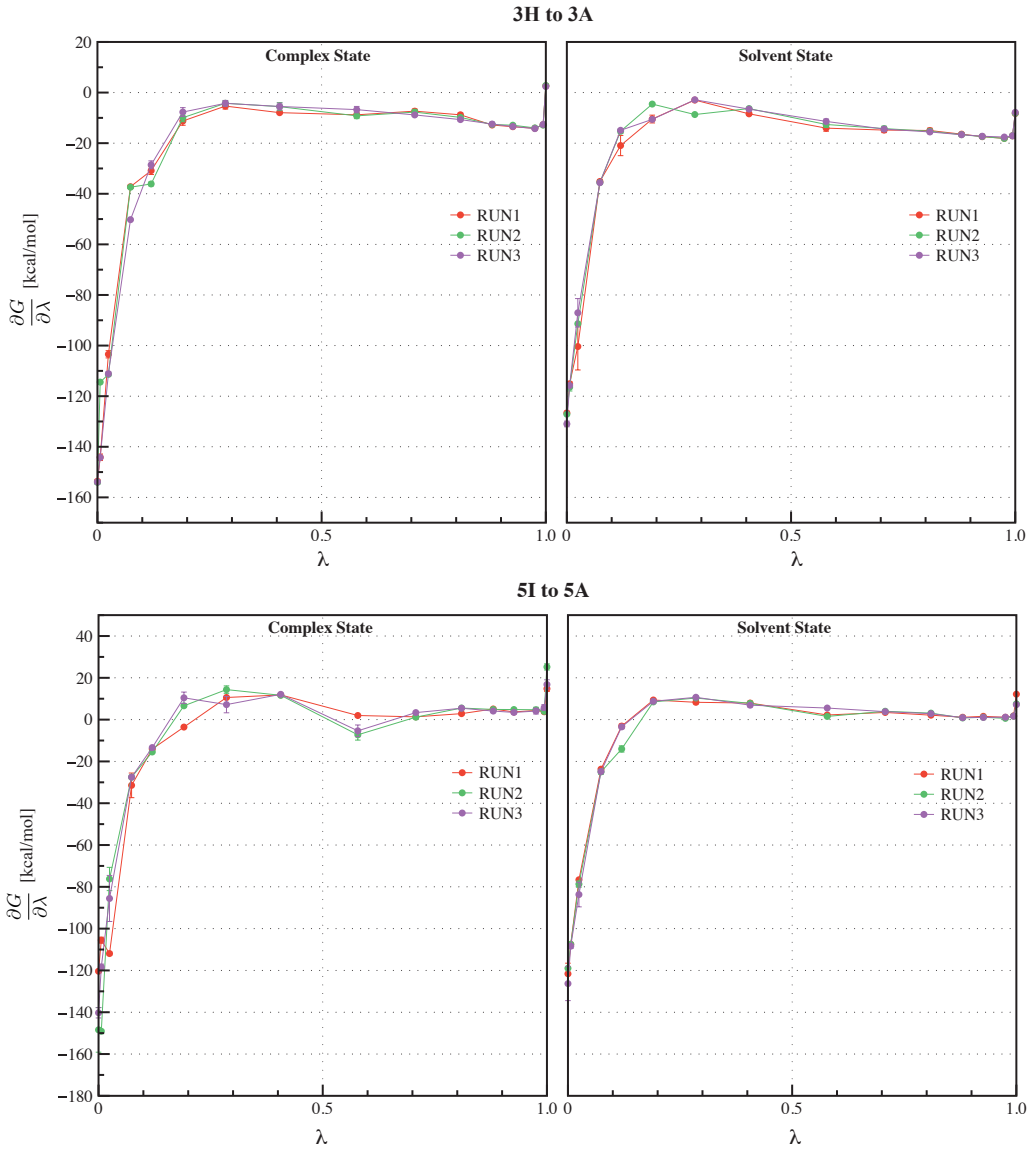


Figure 5.11: *Free energy gradients for the considered alchemical mutations in complex and solvent states recorded over three independent runs. In this case, the protein and the ligands were restrained by applying potentials 5.2, 5.3 while, the ligand-solvent interactions were limited by creating a penalised solvent volume around the ligand by using the potential 5.4. The mutations* $5I \rightarrow 5A$ *in complex state and* $3H \rightarrow 3A$ *in solvent state present convergence issues in many window values and, therefore, the predicted non-additivity is unreliable.*

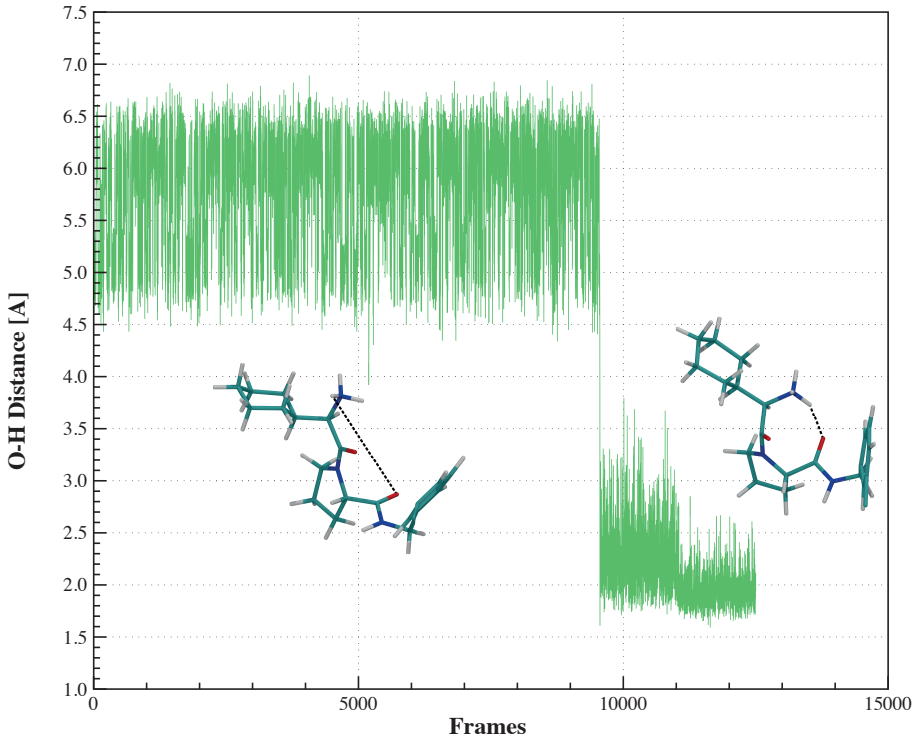


Figure 5.12: *The intra-molecular distance between one of the hydrogen of the amino group in the 5I ligand and one of the indicated ligand oxygen atoms recorded along the complex state simulation* $5I \rightarrow 5A$ *for the window value* $\lambda = 0$ *applying the implemented potentials 5.2, 5.3 and 5.4. The dehydrated volume produced by the potential 5.4 facilitates and intra-molecular interaction changing the binding mode of the ligand along the simulation and producing noisy gradient values.*

summarise and give conclusions on the whole thesis work.

6

Conclusions

In the pharmaceutical industry the use of computational methods has significantly increased in the last 25 years. *In-silico* methods have made meaningful contributions in the drug discovery and development process such as lead generation, lead optimization, prediction of drug likeness, de novo design, ligand docking, modulation of ADME and toxicity. However, the precise determination of the binding affinity and its entropic and enthalpic components is still considered the "Holy Grail" of computational chemists[137]. The main factors that discourage the routine use of molecular simulations to support medicinal chemist workflows are computational cost, inaccuracy and setup difficulties. To be useful, a computational prediction should be performed quicker than the related experimental measurement. In the statistical mechanics framework, the computation of the binding affinity requires the sampling of many configurations of a protein-ligand complex to ensure that the ensemble average produces results comparable with the

experimental measurements. This requirement could be very difficult to achieve for biomolecular systems because of the high number of degree of freedoms involved in the simulations. Inaccuracies in the binding affinity prediction are often correlated to molecular models used to describe biomolecular systems. Force field parameters are often optimised to predict specific properties for specific classes of molecules and they could in principle fail to predict other properties when they are applied on molecular systems that differ from a training set. In addition many computational models used to simulate biomolecular systems neglect important aspects such as polarizability or changes in covalent interactions. Currently, the calculation of binding affinity by molecular simulations seems to be more an "art to master" than a standardised protocol to use. The modern computational chemist needs to be a skilled computer scientist in addition to being a "clever observer", "modeller" and "result analyst" of the underlying physico-chemical processes. The use of many configuration files, terminal applications and operating system knowledge are required and the learning curve might be very steep indeed. The above reasons are just the main problems to overcome before pharmaceutical companies will consider binding affinity calculations by molecular simulations a valuable and effective new way for the development of new drugs.

This thesis aims to address the issue of computational cost of binding affinity calculations by delivering a new tool for the calculation of relative binding affinities, allowing larger compound data sets to be studied in a reasonable time. This was achieved by merging two pieces of software: Sire and the OpenMM APIs. The former is a flexible molecular modelling framework for bio-molecular modelling, enabling easy definition and editing of molecular parameters and implementation of new molecular simulation methods. On the other hand, the OpenMM APIs are able to perform efficient MD simulations on modern specialised hardware such as GPUs. The resulting implementation is flexible and quick enough to computationally explore "new science" which lacks in alternative commercial software packages. Frequently such software packages are restricted to predefined and rigid simulation schemes, which do not fit in the scientific research context where many new ideas need to be tested and validated.

Chapter two described the relative free energy implementation. This is based on alchemical transformations and in particular on the single topology method. This technique uses a shared scaffold to transform a ligand into another. Bonded and non-bonded force field parameters are linearly interpolated between the starting and final ligand force field parameters by using the coupling parameter approach. The single topology method is used in conjunction with the FDTI method to calculate free energy gradients over a selected set of coupling parameter values to be numerically integrated and to calculate the free energy change associated with the ligand transformation. In the implementation, three categories of atoms were defined to perform the mutation: "to dummy", "from dummy" and "hard" atoms. These atom groups respectively represent atoms that can disappear, appear or be consistently present along the alchemical mutation. The presence of "to dummy" and "from dummy" atom groups might produce numerical instabilities along the alchemical simulations and, therefore, the non-bonded interactions were softened in the implementation by using a soft-core potential. The implementation was initially tested on 100 alchemical mutations in vacuum to perform single point energy calculations. Subsequently, relative free energies of solvation of many small molecules were evaluated. The results were in agreement with experimental data.

Chapter 3 investigated from a computational point of view the effect of molecular flexibility on conformational equilibria of a set of molecules with two main moieties linked together by flexible carbon chains. These molecules can experimentally adopt different conformers in chloroform solution and, in particular, they are able to form an intra-molecular hydrogen bond. Experimentally the free energy change between a "folded" state where the intra-molecular hydrogen bond was formed and an "unfolded" state where the hydrogen bond was broken was measured for different lengths of the carbon chain. Computational simulations modelled the experimental setting, trying to reproduce the free energy change between the "folded" and "unfolded" states by using molecular dynamics simulations. In the computational setting a major problem was the derivation of parameters and, in particular, the charge calculations. The AM1-BCC charge

method was not able to accurately reproduce the experimental data and the CM5 charge method was preferred. However, results were not in good agreement with the experimental data. In many cases the source of discrepancy was related to the different sensitivity between the experimental and computational model between the "folded" and "unfolded" populations. In order to improve the results, other force field parameters were changed such as Lennard-Jones and torsional parameters. The results showed that these parameters do not have a significant impact on the computed free energies. The charge calculation seems to have the greatest impact on the simulations, and the use of more advanced models such as polarizable force fields may improve the agreement between predictions and experimental data.

In the lead optimisation stage iterative SAR studies are performed to improve the binding affinity of promising hits. However, this process is particularly difficult. The main causes are related to the optimisation of the entropic and enthalpic contributions. Binding enthalpy optimisation is notoriously difficult to improve and it depends on the optimisation of VdW forces/hydrogen bonds and desolvation of polar groups. On the other hand, entropic optimisations are related to conformational entropy changes and solvation entropy. One of the main simplification introduced in the routine workflow of medicinal chemists is the additivity of the binding affinity i.e. the assumption that free energy can be decomposed into sum of independent components ascribed to specific parts of a system. Chapter four investigated a series of congeneric inhibitors of the Thrombin protein that present non-additivity behaviour. The implemented code was applied to the calculation of the relative binding affinities of two series of Thrombin inhibitors named the 3 and 5 series. Results were in in good agreement with the experimental data in many cases and in order to quantify the level of agreement and in-depth error analysis was conducted between the experimental and the predicted data by using different statistical quantities such as the coefficient of determination (R^2), the mean unsigned error (MUE) and the predictive index (PI). The error analysis showed that the predicted statistical quantities are in the selected error analysis confidence interval and, on average, the experimental and predicted data of the

5 series are quantitatively better than the 3 series. Although the accuracy of the predicted binding affinities is significant, the prediction of the non-additivity levels among the thrombin inhibitors was poor. One of the main issues is the systematic underestimation of the binding affinity in the 3 series compared to the 5 series, which could be related to force field parameterisation problems, and further investigations are required to prove or disprove this point.

Even though the predicted non-additivity levels were poor, in at least one system the predicted level was comparable with the experimental data and, the non-additivity origins were analysed for this system only. In order to address the possible non-additivity sources three hypotheses were suggested i.e. changes in the protein or ligand flexibilities or changes in the ligand-solvent interactions. With the aim of testing these hypotheses, three computational experiments were designed with the goal to suppress the non-additivity present in the system. In a first computational experiment, the protein was rigidified by applying positional restraints to selected protein atoms and results showed that the non-additivity was suppressed from the system. A second computational experiment rigidified the ligands by using distance restraints applied between selected atom pairs, while a final experiment tried to limit the ligand-solvent interactions by defining a penalised solvent volume around the ligands. Unfortunately, difficulties with convergence of the calculated free energies biased these two last computational experiments and, therefore, it was not possible to clearly prove or disprove their impact on the selected system. Although the protein rigidification setup succeeded in suppressing non additivity for just one studied system, this could suggest that changes in the protein flexibility might be the origin of the non-additivity in the Thrombin inhibitors, and this would corroborate with the experimental observation of a binding site rigidification as evidenced by B factor measurements.

An initial goal of this thesis was also to develop a computationally efficient relative free energy implementation; however, benchmarks have not been presented so far. An in-depth analysis of the code speed-up has not been performed. In the end, the research project focused more on result correctness than a pure code optimization. Sire and the OpenMM APIs have been extensively tested and op-

timised and the implementation speed-up relies on these two software packages. Indeed, the ultimate goal was to merge the two pre-existing codes. However, in the code interface particular attention was taken to handle efficiently system creations and trajectories storing to avoid bottlenecks. The solvated Thrombin system presented on average $34 \cdot 10^3$ atoms; it was possible to run relative binding affinity calculations by using two different GPU architectures. Roughly speaking the time required to run 10 ns simulation per window on an NVidia M2090 GPU card was approximately 14 hours while the time was nearly halved by using a more recent GPU architecture such as the NVidia K20. This result is quite encouraging. The same piece of code applied on the same system and tested on two architectures (the former ca. 3 year older than the latter), was able to half the computational time. This suggests that the computational cost is no longer a problem in the binding free energy calculations context and efforts should address other problems. It was not possible to make a comparison between CPU versus GPU benchmarks but it is the author's opinion that this is an unrealistic comparison. Architecturally, the CPU is composed of only few cores with lots of cache memory that can handle a few software threads at a time. It is also designed to handle interrupts, virtual memory and storage, which are required by the modern operating systems to perform most of the processing in everyday computing. In contrast, a GPU is composed of hundreds of cores that can handle thousands of threads simultaneously. Modern GPUs are capable of performing vector operations and floating-point arithmetic, with the latest cards that full support double-precision floating-point arithmetic and make them most suited to highly parallelizable operations such as in scientific computing. However, a comparison should be possible on their effective market cost. Gaming GPUs that can be used to perform scientific calculations have on average the same market cost of a server CPU solution but benefiting of speed-up of 100 times in many scientific computing applications.

The implementation calculates relative binding free energies and not absolute binding free energies. Usually, the computation of a relative free energy of binding is thought to be more efficient compared to an absolute free energy of binding.

This is partly due to the hope that in the hypothesis of incorrect sampling for both ligands the errors in the difference will vanish, but in general, this cancellation of errors cannot be taken for granted. In addition, in the SAR stage there is the need to compare the binding affinity of prominent hits and therefore in drug development and discovery process this quantity is more significant to enhance the medicinal chemist workflow. In the studied Thrombin inhibitors the predicted order potency assessed by using the PI index was quite significant for both the examined series with a value greater than 0.9 ($-1 \leqslant \mathrm{PI} \leqslant 1$).

The implemented code is based on the single topology method where bonded and non-bonded parameter terms are interpolated between a starting a final ligand and by using a coupling parameter technique. On the other hand, in the dual topology method both ligands are present at the same time; the interatomic interactions are gradually scaled between the ligands and the environment by using the coupling parameter such that at the beginning and in the end of the ligand transformation it is respectively present the starting and final ligand only. An advantage of the single topology method is that the perturbation is localized and it should produce faster convergence. On the other hand, the two ligands need to be structurally similar while, in the dual topology they can be arbitrarily different. The soft-core potential is extremely useful to mitigate numerical instabilities in simulations where atoms can appear or disappear. However, its correct setting might be very difficult in some simulations. For instance, in one of the Thrombin inhibitor mutation a monoatomic ion in solution was trapped by one "to dummy" carbon atom. The problem was related to the quicker softening of LJ interactions, compared to the softening of the Coulombic interactions. The selected decoupling schedule gave reasonable results in most of the simulated system, but it is hard to find a general protocol. A possibility is to decouple the LJ and the Coulomb terms in two separate simulations however, this would double the computational time.

A key point of the implementation is its flexibility. During the entire project the code was many time extended to explore new methods, for instance the implementation of distance restraints or the application of a penalized solvent volume

around the ligand in the Chapter five. The definition of interaction groups in Sire and OpenMM is very easy and allows simulations that would be very difficult to implement with other simulation packages. Another advantage of this implementation is that the sampling of a molecular system can be done by using MD and MMC. The hybrid method could gain from both the previous techniques. The size of the time step in the MD method is a significant drawback, which affects its accuracy; large time steps can break the convergence as well as the energy conservation in isolated systems. On the other hand, the movement of all the particles in one integration step is an advantage. MMC is not deterministic and, as a consequence, the generation of new moves can be completely arbitrary; the only restriction is in respect of the detailed balance equation. Hybrid MD-MMC relaxes the restriction on the size of the time step. A high value will produce a "bad movement" in the phase space of MD simulations but if the resulting configuration is subjected to a Metropolis acceptance test the canonical ensemble is preserved. Furthermore, by using MD to generate moves the system tends to move in regions of configuration space with lower energy and hence moves are more "clever" than simple displacements. The implementation of the hybrid code would be not particular difficult at this stage of the project. Sire is able to perform fast MMC moves and OpenMM is able to perform fast MD moves. Therefore, the code already presents an advanced infrastructure where to build an extension to support hybrid MD-MMC. This would be best accomplished by writing Sire objects in C++ that requires adequate programming expertise, but fortunately once written, the Sire object classes are accessible to non-expert users via a Python front end.

In conclusion, the author is confident that the broad aim of this research project has been achieved. An efficient implementation of relative binding affinity calculation protocols by molecular simulations has been developed and made available to the scientific community. The software has been extensively tested by applications on different biomolecular systems, such as the Thrombin enzyme, showing good agreement with experimental data even though the origins of non-additivity effects in this system needs further investigations. In general, software

tools should be used to help researchers to test new ideas and simplify their routinely work. There have already been examples where the code was used not just in the hands of the thesis author. The code was involved in the validation of an implementation of the Grid-Cell theory resulting in a publication[138]. In addition, recently, the implementation was extended to support the GLYCAM force field and, binding free energy calculations of protein-carbohydrate complexes were performed producing good agreement with experimental data[139]. To summarise, the implementation is flexible enough to test and validate new ideas in free energy calculations and overall in molecular simulations.

Bibliography

[1] Janet Woodcock and Raymond Woosley. The FDA critical path initiative and its influence on new drug development. *Annual Review of Medicine*, 59 (1):1–12, 2008.

[2] P.A.M Dirac. Quantum mechanics of many-electron systems. In *Quantum Mechanics of Many-Electron Systems*, volume 123 of *A, Containing Papers of a Mathematical and Physical Character*, pages 714–733. Royal Society of London, 1929.

[3] A. M. Turing. On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, s2-42(1):230–265, 1937.

[4] John von Newmann. First draft of a report on the edvac. Technical report, University of Pennsylvania Moore School of Electrical Engineering, 1945.

[5] Haile J. M. *Molecular Dynamics Simulation: Elementary Methods*. Wiley-Interscience, New York, 1997.

[6] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[7] Enrico Fermi, G.J. Pasta, and S Ulam. Studies of non linear problems. Technical Report LA-1940, Los Alamos Laboratories, 1955.

[8] B. J. Alder and T. E. Wainwright. Studies in molecular dynamics. i. general method. *The Journal of Chemical Physics*, 31(2):459–466, 1959.

[9] Winau Floria, Westphal Otto, and Winau Rolf. Paul ehrlich - in search of the magic bullet. *Microbes and Infection*, 6:786–789, 2004.

[10] Christopher Adams and Van Brantner. Spending on new drug development. *Health Economics*, 19(2):130–141, 2010.

[11] Frank Sams-Dodd. Drug discovery: selecting the optimal approach. *Drug Discovery Today*, 11(9-10):465–472, 2006.

[12] Paweł Szymański, Magdalena Markowicz, and Elżbieta Mikiciuk-Olasik. Adaptation of High-Throughput Screening in Drug Discovery—Toxicological Screening Tests. *International Journal of Molecular Sciences*, 13(1):427–452, 2011.

[13] Ernesto Freire. Do enthalpy and entropy distinguish first in class from best in class? *Drug Discovery Today*, 13(19-20):869–874, 2008.

[14] Charles L. Brooks. *Proteins: a theoretical perspective of dynamics, structure, and thermodynamics.* Number v. 71 in Advances in chemical physics. J. Wiley, New York, 1988.

[15] David Mobley and Pavel Klimovich. Perspective: Alchemical free energy calculations for drug discovery. *The Journal of Chemical Physics*, 137(23): 230901, 2012.

[16] M.Mertz Kenneth, Ridge Dragmar, and H. Reynolds Charles. *Drug Design. Stucture- and Ligand-based Approaches.* Cambridge University Press, 2010.

[17] Philip J. Hajduk and Daryl R. Sauer. Statistical Analysis of the Effects of Common Chemical Substituents on Ligand Potency. *Journal of Medicinal Chemistry*, 51(3):553–564, 2008.

[18] Raimund Mannhold, Hugo Kubinyi, and Gerd Folkers. *Hit and Lead Profiling: Identification and Optimization of Drug-like Molecules.* Wiley-VCH, 1 edition edition, September 2009.

[19] David W. Borhani and David E. Shaw. The future of molecular dynamics simulations in drug discovery. *Journal of Computer-Aided Molecular Design*, 26(1):15–26, 2012.

[20] Julie R. Schames, Richard H. Henchman, Jay S. Siegel, Christoph A. Sotriffer, Haihong Ni, and J. Andrew McCammon. Discovery of a novel binding trench in HIV integrase. *Journal of Medicinal Chemistry*, 47(8):1879–1881, 2004.

[21] Jacob D. Durrant and J. Andrew McCammon. Molecular dynamics simulations and drug discovery. *BMC biology*, 9:71, 2011.

[22] Ettore Majorana and R. N. Mantegna. The value of statistical laws in physics and social sciences. In Giuseppe Franco Bassani and Council of the Italian Physical Society, editors, *Ettore Majorana Scientific Papers*, pages 237–260. Springer Berlin Heidelberg, 2006.

[23] Mark E. Tukerman. *Statistical Mechanics: Theory and Molecular Simulation.* Oxford University Press, 2010.

[24] J. Willard Gibbs. *Elementary principles in statistical mechanics : developed with especial reference to the rational foundation of thermodynamics.* New York : C. Scribner, 1902.

[25] M. Born and R. Oppenheimer. Zur quantentheorie der molekeln. *Annalen der Physik*, 389(20):457–484, 1927.

[26] Michael Griebel, Stephan Knapek, and Gerhard Zumbusch. *Numerical Simulation in Molecular Dynamics: Numerics, Algorithms, Parallelization, Applications.* Springer Publishing Company, Incorporated, 2007.

[27] R. S. Mulliken. Electronic population analysis on LCAO–MO molecular wave functions. i. *The Journal of Chemical Physics*, 23(10):1833–1840, 1955.

[28] Araz Jakalian, Bruce L. Bush, David B. Jack, and Christopher I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. method. *Journal of Computational Chemistry*, 21(2):132–146, 2000.

[29] René Sperb Ilario G. Tironi. A generalized reaction field method for molecular dynamics simulations. *The Journal of Chemical Physics*, 102(13):5451–5459, 1995.

[30] Y. Toukmaji Abdulnour and A. Board John. Ewald summation techniques in perspective: a survey. *Computer Physics Communications*, pages 73–92, 1996.

[31] Andrew Leach. *Molecular Modelling: Principles and Applications*. Prentice Hall, Harlow, England ; New York, 2 edition edition, 2001.

[32] Norman L. Allinger. Conformational analysis. 130. MM2. a hydrocarbon force field utilizing v1 and v2 torsional terms. *Journal of the American Chemical Society*, 99(25):8127–8134, 1977.

[33] Wendy Cornell, Piotr Cieplak, Christopher Bayly, Ian Gould, Kenneth Merz, David Ferguson, David Spellmeyer, Thomas Fox, James Caldwell, and Peter Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, 117(19): 5179–5197, 1995.

[34] A. D. MacKerell, C. L. Brooks, L. Nilsson, B. Roux, Y. Won, and M. Karplus. *CHARMM: The Energy Function and Its Parameterization with an Overview of the Program*, volume 1 of *The Encyclopedia of Computational Chemistry*, pages 271–277. John Wiley & Sons, 1998.

[35] W. F. van Gunsteren and Berendsen. Groningen Molecular Simulation (GROMOS) Library Manual. Biomos, Nijenborgh 16, Groningen, NL, 1987.

[36] William L. Jorgensen and Julian Tirado-Rives. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*, 110(6):1657–1666, 1988.

[37] Thomas A Halgren. Potential energy functions. *Current Opinion in Structural Biology*, 5(2):205–210, 1995.

[38] Martin Field. *A Practical Introduction to the Simulation of Molecular Systems*. Cambridge University Press, 2 edition, 2007.

[39] Christopher J. Cramer. *Essentials of Computational Chemistry: Theories and Models*. Wiley, 2 edition edition, 2004.

[40] Norman L. Allinger, Young H. Yuh, and Jenn Huei Lii. Molecular mechanics. The MM3 force field for hydrocarbons. 1. *Journal of the American Chemical Society*, 111(23):8551–8566, 1989.

[41] Norman L. Allinger, Kuohsiang Chen, and Jenn-Huei Lii. An improved force field (MM4) for saturated hydrocarbons. *Journal of Computational Chemistry*, 17(5-6):642–668, 1996.

[42] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society*, 114(25): 10024–10035, 1992.

[43] Christophe Chipot. *Free Energy Calculations: Theory and Applications in Chemistry and Biology (Springer Series in Chemical Physics)*. Springer, 2007.

[44] John Kirkwood. Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics*, 3(5):300–313, 1935.

[45] Robert Zwanzig. High-temperature equation of state by a perturbation method. i. nonpolar gases. *The Journal of Chemical Physics*, 22(8):1420–1426, 1954.

[46] David Chandler Lawrence R. Pratt. Theory of the Hydrophobic Effect. *The Journal of Chemical Physics*, 67(8):3683–3704, 1977.

[47] I. R. McDonald and K. Singer. Machine Calculation of Thermodynamic Properties of a Simple Fluid at Supercritical Temperatures. *Journal of Chemical Physics*, 47:4766–4772, 1967.

[48] Jong K. Lee, J. A. Barker, and Farid F. Abraham. Theory and Monte Carlo simulation of physical clusters in the imperfect vapor. *The Journal of Chemical Physics*, 58(8):3166–3180, 1973.

[49] Michael R. Mruzik, Farid F. Abraham, Donald E. Schreiber, and G. M. Pound. A Monte Carlo study of ion–water clusters. *The Journal of Chemical Physics*, 64(2):481–491, 1976.

[50] Susumu Okazaki, Koichiro Nakanishi, Hidekazu Touhara, and Yoshinori Adachi. Monte Carlo studies on the hydrophobic hydration in dilute aqueous solutions of nonpolar molecules. *The Journal of Chemical Physics*, 71(6): 2421–2429, 1979.

[51] C. Pangali, M. Rao, and B. J. Berne. A Monte Carlo simulation of the hydrophobic interaction. *The Journal of Chemical Physics*, 71(7):2975–2981, 1979.

[52] Bhalachandra L. Tembre and J. Andrew Mc Cammon. Ligand-receptor interactions. *Computers & Chemistry*, 8(4):281 – 283, 1984.

[53] William Jorgensen and C Ravimohan. Monte carlo simulation of differences in free energies of hydration. *The Journal of Chemical Physics*, 83(6):3050–3054, 1985.

[54] Jayaraman Chandrasekhar, Scott F. Smith, and William L. Jorgensen. SN2 reaction profiles in the gas phase and aqueous solution. *Journal of the American Chemical Society*, 106(10):3049–3050, May 1984.

[55] P. A. Bash, U. C. Singh, R. Langridge, and P. A. Kollman. Free Energy Calculations by Computer Simulation. *Science*, 236(4801):pp. 564–568, 1987.

[56] P. A. Bash, U. C. Singh, F. K. Brown, R. Langridge, and P. A. Kollman. Calculation of the relative change in binding free energy of a protein-inhibitor complex. *Science*, 235(4788):574–576, 1987.

[57] Shashidhar N. Rao, U. Chandra Singh, Paul A. Bash, and Peter A. Kollman. Free energy perturbation calculations on binding and catalysis after mutating Asn 155 in subtilisin. *Nature*, 328(6130):551–554, 1987.

[58] C. Jarzynski. Nonequilibrium Equality for Free Energy Differences. *Physical Review Letters*, 78(14):2690–2693, 1997.

[59] Stephen H. Fleischman and Charles L. Brooks Iii. Thermodynamics of aqueous solvation: Solution properties of alcohols and alkanes. *The Journal of Chemical Physics*, 87(5):3029–3037, 1987.

[60] T. P. Straatsma and H. J. C. Berendsen. Free energy of ionic hydration: Analysis of a thermodynamic integration technique to evaluate free energy differences by molecular dynamics simulations. *The Journal of Chemical Physics*, 89(9):5876–5886, 1988.

[61] Thomas C. Beutler, Alan E. Mark, René C. van Schaik, Paul R. Gerber, and Wilfred F. van Gunsteren. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chemical Physics Letters*, 222(6):529–539, 1994.

[62] Gerhard Hummer, Lawrence R. Pratt, and Angel E. García. Free Energy of Ionic Hydration. *The Journal of Physical Chemistry*, 100(4):1206–1215, 1996.

[63] David A. Pearlman and Peter A. Kollman. The overlooked bond-stretching contribution in free eenrgy perturbation calculations. *The Journal of Chemical Physics*, 94(6):4532–4545, 1991.

[64] Stefan Boresch and Martin Karplus. The Jacobian factor in free energy simulations. *The Journal of Chemical Physics*, 105(12):5145–5154, 1996.

[65] Erin M. Duffy and William L. Jorgensen. Prediction of Properties from Simulations: Free Energies of Solvation in Hecadecane, Octanol and Water. *Journal of the American Chemical Society*, 122(12):2878–2888, 2000.

[66] William L. Jorgensen, Juliana Ruiz-Caro, Julian Tirado-Rives, Aravind Basavapathruni, Karen S. Anderson, and Andrew D. Hamilton. Computer-aided design of non-nucleoside inhibitors of HIV-1 reverse transcriptase. *Bioorganic & Medicinal Chemistry Letters*, 16(3):663–667, 2006.

[67] David A. Pearlman and Paul S. Charifson. Are Free Energy Calculations Useful in Practice? A Comparison with Rapid Scoring Functions for the p38 MAP Kinase Protein SystemâĂă. *Journal of Medicinal Chemistry*, 44 (21):3417–3423, October 2001.

[68] Thomas Simonson, Georgios Archontis, and Martin Karplus. Continuum Treatment of Long-Range Interactions in Free Energy Calculations. Application to Protein-Ligand Binding. *The Journal of Physical Chemistry B*, 101(41):8349–8362, October 1997.

[69] Jessica M. J. Swanson, Richard H. Henchman, and J. Andrew McCammon. Revisiting Free Energy Calculations: A Theoretical Connection to MM/PBSA and Direct Calculation of the Association Free Energy. *Biophysical Journal*, 86(1):67–74, 2004.

[70] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002.

[71] Alec Hodel, Thomas Simonson, Robert O. Fox, and Axel T. Brunger. Conformational substrates and uncertainty in macromolecular free energy calculations. *The Journal of Physical Chemistry*, 97(13):3409–3417, 1993.

[72] Behrooz Parhami. *Introduction to Parallel Processing: Algorithms and Architectures*. Kluwer Academic Publishers, 1999.

[73] David B. Kirk and Wen-mei W. Hwu. *Programming Massively Parallel Processors: A Hands-on Approach*. Morgan Kaufmann Publishers Inc., 2010.

[74] Christopher J. Woods, Julien Michel, and Gaetano Calabro. Sire, 2014. URL `http://siremol.org/Sire/Home.html`.

[75] P. Eastman and V.S. Pande. OpenMM: A Hardware-Independent Framework for Molecular Simulations. *Computing in Science Engineering*, 12(4): 34–39, 2010.

[76] Bernhard Baum, Laveena Muley, Michael Smolinski, Andreas Heine, David Hangauer, and Gerhard Klebe. Non-additivity of functional group contributions in protein–ligand binding: A comprehensive study by crystallography and isothermal titration calorimetry. *Journal of Molecular Biology*, 397(4): 1042–1054, 2010.

[77] Christine Peter, Chris Oostenbrink, Arthur van Dorp, and Wilfred van Gunsteren. Estimating entropies from molecular dynamics simulations. *The Journal of Chemical Physics*, 120(6):2652–2661, 2004.

[78] Loup Verlet. Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Physical Review*, 159(1):98–103, 1967.

[79] William C. Swope, Hans C. Andersen, Peter H. Berens, and Kent R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics*, 76(1):637–649, 1982.

[80] L. V. Woodcock. Isothermal molecular dynamics calculations for liquid salts. *Chemical Physics Letters*, 10(3):257–261, 1971.

[81] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684–3690, 1984.

[82] Hans C. Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of Chemical Physics*, 72(4):2384–2393, 1980.

[83] Shūichi Nosé. A molecular dynamics method for simulations in the canonical ensemble. *Molecular Physics*, 52(2):255–268, 1984.

[84] J. G. Kirkwood and B. J Alder. *Theory of Liquids*. Gordon and Breach, 1968.

[85] Stefan Boresch, Franz Tettinger, Martin Leitgeb, and Martin Karplus. Absolute binding free energies: a qualitative approach to their calculation. *The Journal of Physical Chemistry B*, 107(35):9535–9551, 2003.

[86] David L. Mobley, John D. Chodera, and Ken A. Dill. On the use of orientational restraints and symmetry corrections in alchemical free energy calculations. *The Journal of Chemical Physics*, 125(8):084902, 2006.

[87] David A. Pearlman. A comparison of alternative approaches to free energy calculations. *The Journal of Physical Chemistry*, 98(5):1787–1493, 1994.

[88] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, Cambridge, UK ; New York, 3 edition edition, 2007.

[89] J. Michel, ML Verdonk, and JW Essex. Protein-ligand complexes: Computation of the relative free energy of different scaffolds and binding modes. *J. Chem. Theory Comput.*, 3:1645–1655, 2007.

[90] J.A. Barker and R.O. Watts. Monte Carlo studies of the dielectric properties of water-like models. *Molecular Physics*, 26(3):789–792, 1973.

[91] R.O. Watts. Monte Carlo studies of liquid water. *Molecular Physics*, 28(4): 1069–1083, 1974.

[92] Stefan Boresch and Martin Karplus. The role of bonded terms in free energy simulations: 1. theoretical analysis. *The Journal of Physical Chemistry A*, 103(1):103–118, 1999.

[93] James J. McGregor. Backtrack search algorithms and the maximal common subgraph problem. *Software: Practice and Experience*, 12(1):23–34, 1982.

[94] Solomon W. Golomb and Leonard D. Baumert. Backtrack Programming. *J. ACM*, 12(4):516–524, 1965.

[95] D. J. Edwards and T. P. Hart. The Alpha-Beta Heuristic. Memo 30, MIT Compuation Center, 1961.

[96] Hannes Loeffler, Christopher J. Woods, and Julien Michel. FeSetup, 2014. URL `http://ccpforge.cse.rl.ac.uk/gf/project/ccpbiosim/`.

[97] Junmei Wang, Romain M. Wolf, James W. Caldwell, Peter A. Kollman, and David A. Case. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25(9):1157–1174, 2004.

[98] Junmei Wang, Wei Wang, Peter A. Kollman, and David A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling*, 25(2):247–260, 2006.

[99] Peter Easteman Pande V. *OpenMM Users Manual and Theory Guide*. Simtk, 6.1 edition, 2014.

[100] Conrad Shyu and Marty Ytreberg. Reducing the bias and uncertainty of free energy estimates by using regression to fit thermodynamic integration data. *Journal of Computational Chemistry*, 30(14):2297–2304, 2009.

[101] Kim-Hung Chow and David M. Ferguson. Isothermal-isobaric molecular dynamics simulations with Monte Carlo volume sampling. *Computer Physics Communications*, 91(1–3):283–289, 1995.

[102] Johan Åqvist, Petra Wennerström, Martin Nervall, Sinisa Bjelic, and Bjørn O. Brandsdal. Molecular dynamics simulations of water and biomolecules with a monte carlo constant pressure algorithm. *Chemical Physics Letters*, 384(4–6):288–294, 2004.

[103] David L. Mobley, Christopher I. Bayly, Matthew D. Cooper, Michael R. Shirts, and Ken A. Dill. Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations. *Journal of Chemical Theory and Computation*, 5(2):350–358, 2009.

[104] Reinhard W. Hoffmann. Flexible Molecules with Defined Shape—Conformational Design. *Angewandte Chemie International Edition in English*, 31(9):1124–1134, 1992.

[105] M. J. T. Robinson. Conformational equilibria involving substituents lacking conical symmetry. *Pure and Applied Chemistry*, 25(3), 1969.

[106] Aleksandr V. Marenich, Steven V. Jerome, Christopher J. Cramer, and Donald G. Truhlar. Charge Model 5: An Extension of Hirshfeld Population Analysis for the Accurate Description of Molecular Interactions in Gaseous and Condensed Phases. *Journal of Chemical Theory and Computation*, 8 (2):527–541, 2012.

[107] Michael J. S. Dewar, Eve G. Zoebisch, Eamonn F. Healy, and James J. P. Stewart. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. [Erratum to document cited in CA103(2):11627f]. *Journal of the American Chemical Society*, 115(12):5348–5348, 1993.

[108] F. L. Hirshfeld. Bonded-atom fragments for describing molecular charge densities. *Theoretica chimica acta*, 44(2):129–138, 1977.

[109] Schrödinger. Suite 2011: Maestro, version 9.2, 2011.

[110] Wang Junmei, M. Wolf Romain, W. Caldwell James, Kollman Peter A., and Case David A. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 26(1):114–114, 2005.

[111] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov,

R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, N. J. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Oe. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox. Gaussian 09, Revision A.1, 2009.

[112] Aleksandr V. Marenich, Christopher Cramer, and Donald Truhlar. CM5pac, 2013.

[113] Ulrich Essmann, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee, and Lee G. Pedersen. A smooth particle mesh Ewald method. *The Journal of Chemical Physics*, 103(19):8577–8593, 1995.

[114] Piotr Cieplak, James Caldwell, and Peter Kollman. Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: aqueous solution free energies of methanol and N-methyl acetamide, nucleic acid base, and amide hydrogen bonding and chloroform/water partition coefficients of the nucleic acid bases. *Journal of Computational Chemistry*, 22(10):1048–1057, 2001.

[115] Oleg Trott and Arthur J. Olson. AutoDock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, 2010.

[116] Hongjian Li, Kwong-Sak Leung, and Man-Hon Wong. idock: A multithreaded virtual screening tool for flexible ligand docking. In *2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 77–84, 2012.

[117] Gregory L. Warren, C. Webster Andrews, Anna-Maria Capelli, Brian Clarke, Judith LaLonde, Millard H. Lambert, Mika Lindvall, Neysa Nevins, Simon F. Semus, Stefan Senger, Giovanna Tedesco, Ian D. Wall, James M.

Woolven, Catherine E. Peishoff, and Martha S. Head. A Critical Assessment of Docking Programs and Scoring Functions. *Journal of Medicinal Chemistry*, 49(20):5912–5931, 2006.

[118] Kanin Wichapong, Marc Lindner, Somsak Pianwanit, Sirirat Kokpol, and Wolfgang Sippl. Receptor-based 3d-QSAR studies of checkpoint Wee1 kinase inhibitors. *European Journal of Medicinal Chemistry*, 44(4):1383 – 1395, 2009.

[119] A. Mark. Decomposition of the free energy of a system in terms of specific interactions implications for theoretical and experimental studies. *Journal of Molecular Biology*, 240(2):167–176, 1994.

[120] Ken A. Dill. Additivity Principles in Biochemistry. *Journal of Biological Chemistry*, 272(2):701–704, 1997.

[121] Yogendra Patel, Valerie J. Gillet, Trevor Howe, Joaquin Pastor, Julen Oyarzabal, and Peter Willett. Assessment of Additive/Nonadditive Effects in Structure-Activity Relationships: Implications for Iterative Drug Design. *Journal of Medicinal Chemistry*, 51(23):7552–7562, 2008.

[122] Neil D. Rawlings, Dominic P. Tolle, and Alan J. Barrett. MEROPS: the peptidase database. *Nucleic Acids Research*, 32(suppl 1):D160–D164, 2004.

[123] Wolfram Bode. Structure and interaction modes of thrombin. *Blood Cells, Molecules, and Diseases*, 36(2):122 – 130, 2006.

[124] Enrico Di Cera. Thrombin. *Molecular Aspects of Medicine*, 29(4):203 – 254, 2008.

[125] Charles T. Esmon. The Protein C Pathway. *Chest*, 124(3_suppl):26S–32S, 2003.

[126] Kenneth G. Mann. Thrombin Formation. *Chest*, 124(3_suppl):4S–10S, 2003.

[127] Youhna Ayala and Enrico Di Cera. Molecular Recognition by Thrombin. Role of the Slows → Fast Transition, Site-specific Ion Binding Energetics and

Thermodynamic Mapping of Structural Components. *Journal of Molecular Biology*, 235(2):733 – 746, 1994.

[128] P Thiagarajan and As Narayanan. Thrombin. In *eLS*. John Wiley & Sons, Ltd, 2001.

[129] Mary Lynn Nierodzik and Simon Karpatkin. Thrombin induces tumor growth, metastasis, and angiogenesis: Evidence for a thrombin-regulated dormant tumor phenotype. *Cancer Cell*, 10(5):355–362, 2006.

[130] Daisuke Mitomo, Yoshifumi Fukunishi, Junichi Higo, and Haruki Nakamura. Calculation of protein ligand binding free energy using smooth reaction path generation (SRPG) method: a comparison of the explicit water model, GB/SA model and docking score function. In *Genome Informatics 2009*, pages 85–97. World Scientific, 2009.

[131] Junsu Ko, Hahnbeom Park Dongseon Lee, Julian Lee Evangelos Coutsias, and Chaok Seok. The FALC-loop web server for protein loop modeling. *Nucleic Acids Research*, 39(2):210–214, 2011.

[132] Christopher J. Woods, Maturos Malaisree, Julien Michel, Ben Long, Simon McIntosh-Smith, and Adrian J. Mulholland. Rapid decomposition and visualisation of protein–ligand binding free energies by residue and by water. *Faraday Discussions*, 169(0):477–499, 2014.

[133] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins*, 65 (3):712–725, 2006.

[134] T.P. Straatsama M. Zacharias and J. A. McCammon. Separation-shifted scaling, a new scaling method for lennard-jones interactions in thermodynamic integration. *Journal of Chemical Physics*, 12(100), 1994.

[135] Renxiao Wang, Luhua Lai, and Shaomeng Wang. Further development and validation of empirical scoring functions for structure-based binding

affinity prediction. *Journal of Computer-Aided Molecular Design*, 16(1): 11–26, 2002.

[136] Myong-Ho Chae, Florian Krull, Stephan Lorenzen, and Ernst-Walter Knapp. Predicting protein complex geometries with a neural network. *Proteins*, 78(4):1026–1039, 2010.

[137] J E DeLorbe, J H Clements, M G Teresk, A P Benfield, H R Plake, L E Millspaugh, and S F Martin. Thermodynamic and structural effects of conformational constraints in protein-ligand interactions. entropic paradoxy associated with ligand preorganization. *J Am Chem Soc*, 131(46):16758–16770, 2009.

[138] Georgios Gerogiokas, Gaetano Calabró, Richard H. Henchman, Michelle W. Y. Southey, Richard J. Law, and Julien Michel. Prediction of Small Molecule Hydration Thermodynamics with Grid Cell Theory. *Journal of Chemical Theory and Computation*, 10(1):35–48, 2014.

[139] Sushil K. Mishra, Gaetano Calabró, Hannes H. Loeffler, Julien Michel, and Jaroslav Koča. Evaluation of Selected Classical Force Fields for Alchemical Binding Free Energy Calculations of Protein-Carbohydrate Complexes. *Journal of Chemical Theory and Computation*, 11(7):3333–3345, 2015.